2-20-2015

G-Quadruplex Dna Structures And Site Specific Genetic Instability

Jonathan David Williams
Illinois State University, jdwill2@ilstu.edu

G-QUADRUPLEX DNA STRUCTURES AND SITE SPECIFIC GENETIC INSTABILITY

Jonathan D. Williams

166 Pages May 2015

Repetitive DNA comprises a majority of the human genome yet functions and overall impacts on site-specific genetic instability are not fully defined. Repetitive G-rich sequences have the propensity to form G-quadruplex (G4), which are stable non-B form DNA structures. G4 structures are conspicuously found at regions of site-specific instability. Even so, human genomic loci capable of forming this structure and their connection to DNA rearrangements are just beginning to be elucidated. My dissertation focuses on G4 structures and their capacity to promote site-specific changes in the human genome, particularly at oncogenes. I identified and investigated new biologically relevant G4 loci in the human genome, using novel computational approaches. The ability for G4 structure formation at subsequent G4 loci was assayed *in vitro* using multiple complimentary techniques. Using human sequence variation databases, these loci showed evidence of increased mutagenesis on both the small and large-scale. At the experimental level, I focused on the frequently translocated oncogene *TCF3* and connected its instability with G4 structure formation. Finally,

I examined how factors functioning in a highly conserved repair pathway, mismatch repair, respond to G4 DNA. My results provide new insights into site-specific genetic instability at repetitive guanine sequences, and offer a new perspective on the biological impact of G4 structure formation.

G-QUADRUPLEX DNA STRUCTURES AND SITE SPECIFIC GENETIC INSTABILITY

JONATHAN D. WILLIAMS

A Dissertation Submitted in Partial
Fulfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

School of Biological Sciences

ILLINOIS STATE UNIVERSITY

2015

G-QUADRUPLEX DNA STRUCTURES AND SITE SPECIFIC GENOME INSTABILITY

JONATHAN D. WILLIAMS

COMMITTEE MEMBERS:

Erik D. Larson, Chair

Tom Hammond

Wade Nichols

John Sedbrook

Brian Wilkinson

ACKNOWLEDGMENTS

I would especially like to thank my PI Dr. Erik Larson for providing me the opportunity to receive my PhD and putting up with me for five years. Also, thank you to my committee of Dr. Hammond, Dr. Nichols, Dr. Sedbrook, and Dr. Wilkinson for advice and allowing me to get this far in the program. Dr. Brad Johnson and Dr. Glen Borchert for training me. All the Larson Lab members past and current, especially the undergraduates that worked tirelessly for the labs research. My medical team has also been influential in my success and includes Dr. Josh Smith and Dr. Susan Rausch. Finally, my Mom Margie, Dad Jim, Sister Jenny, and most importantly my Fiancée Samantha, have been the most amazing support system imaginable and I owe all my past and future successes to their love and encouragement. Redbirds for life!

J.D.W.

CONTENTS

TABLES

FIGURES

CHAPTER I

INTRODUCTION, G4 DNA AND GENOME INSTABILITY

Genome instability is a broad term used to describe genetic alterations. These changes can include single nucleotide polymorphisms (SNP), nucleotide insertions or deletions (indels), gene copy number variations (CNV), or aneuploidy. Genome instability arises from many sources, such as DNA synthesis errors, DNA damage, chromosomal rearrangements, inhibition of repair, and the instability of repetitive DNA (Negrini et al., 2010; Kennedy et al., 2012). Genome instability is a concern because it predisposes cells to cancer and complicates treatment (Loeb et al., 2011), but it also characterizes certain loci involved in degenerative neurological disorders (McMurray, 2010). It is important to understand the underpinnings of genome instability because it reveals the molecular causes of human disease. Such information could help improve diagnostics or treatments. One important and poorly defined source of genome instability is repetitive DNA, and this is the primary focus of my dissertation.

There is evidence that repetitive DNA is unstable, and leads to disease. The rearrangements of chromosomes, called translocations, are amongst the most common form of genomic instability leading to sporadic cancer (Bunting, 2013). They occur between specific loci, and the reasons for the rearrangements are generally undefined (Negrini et al., 2010; Kennedy et al., 2012). Deletions, insertions, and inversions are also common large-scale mutagenesis events that lead to human disease (Pikor et al., 2013). Of particular interest are results from computational analyses using existing genome databases, which identify repetitive DNA sequences located near frequent rearrangement breakpoints in

human disease (Katapadi et al., 2012; Wells, 2007; Bacolla et al., 2006; Abeysinghe et al., 2003; Stenson et al., 2003). This implies these repeats play a role in site-specific genetic instability. DNA repeats can adopt a variety of non-B form conformations (Figure 1). Even though these are generally considered to be deleterious, and a major topic of this dissertation, DNA structures could also be important driving forces in evolution (Zhao et al., 2010).

The formation of non-B form DNA structures by repeat sequences most likely promote DNA breaks by interfering with replication or transcription (van Kregten and Tijsterman, 2014; Koole et al., 2014; Yadav et al., 2014). This is supported by multiple experimental systems showing a connection between DNA breaks and structure stability. For instance, when structure-forming human minisatellites were inserted into the end of chromosome 5 in yeast, the rate for chromosome arm loss was increased over 1500-fold (Piazza et al., 2012). Further experimentation using different, but related, repeat sequences discovered that chromosome instability was increased in conditions that promoted structure resolution by helicases or stabilized structure formation (Ribeyre et al., 2009; Piazza 2010, 2012; Yadav et al., 2014). Similar model systems have also shown that DNA structures can inhibit break repair resulting in site-specific instability (van Kregten and Tijsterman, 2014). Therefore, DNA structures actively promote instability and my dissertation stems from these earlier reports.

There are many DNA structures that can potentially form in a cell. This includes hairpins (cruciform), triplex, and G-quadruplex (G4) (Figure 1). The type

of DNA structure that forms is largely dictated by the sequence. For instance, inverted repeats form stable hairpins because of complementary base pairing within a single DNA strand, where as trinucleotide repeats form imperfect, and therefore less stable, hairpins (Figure 2B). G-rich sequences can also form a wide variety of four stranded structures, called G4. Formation of G4 is based on the density and spacing of guanine repeats (Figure 2A). Similar to hairpins, G4 structures have variable levels of stability based on sequence composition (Sen and Gilbert, 1990). However, the precise type of G4 structure that can form from a given sequence is not well understood.

The results described in this dissertation will focus on G4 DNA. Guanine, unlike the other three bases, can form stable base pair interactions with other guanines. This was first reported in the 1960's when researchers noticed that solutions of guanine left on a bench top aggregated (Ralph et al., 1962). It wasn't until decades later that researchers began to characterize the ability of guanine repeats to fold into four-stranded conformations (Sen and Gilbert, 1988, 1990, 1992; Gellert et al., 1962). G4 DNA formation requires repeat sequences of guanine, where Hoogsteen base pairing interactions between guanine bases results in a square, planar, structure that stack upon one another to form a four-stranded structure (Figure 2A) (Gellert et al., 1962; Sen and Gilbert, 1988, 1990). Sequences that can support G4 DNA are variable, but do depend upon the sequence composition (Maizels, 2006, 2012). At a minimum, intra-molecular (or monomolecular) G4 DNA requires the following motif: nGGGnGGGnGGGnGGGn. This is where G represents guanine and n denotes

one or multiple non-guanine bases. The bases separating guanine repeats can vary greatly (n=1-24) and the number of tandem guanines can go well beyond the minimal 3 bases (Burge et al., 2006) (Figure 2A). Using this basic definition in search algorithms, over 300,000 different motifs in the human genome are capable of adopting intra-molecular G4 conformations (Huppert and Balasubramanian, 2005; Todd et al., 2005). Considering that so many potential G4 sequences exist in humans, and the apparent connection between G4 DNA and instability, it is important to characterize the relationship between G4 DNA and DNA damage.

Currently, there is a great deal of evidence that G4 DNA forms *in vivo*. First, there are proteins that specifically bind or metabolize G4 DNA, such as BLM, FANCJ, WRN, Nucleolin, MutSα, PARP, Ku, and RNP-A1 which all have high affinities for G4 structure binding or resolution (Balasubramanian and Neidle, 2009; Cogoi et al., 2008; Sun et al., 1998; Wu et al., 2008; Fry and Loeb, 1999; Larson et al., 2005; González et al., 2009). Second, regions of programmed recombination like immunoglobulin (Ig) loci, likely involve the formation of G4. Transcription of Ig switch regions are necessary for recombination of Ig constant regions, and these sequences readily adopted co-transcriptional G4 DNA *in vitro* (Duquette et al., 2004). Third, G4 specific probes have directly visualized the presence of G4 structures at guanine rich loci in the cell (Biffi et al., 2013). One such region that is extensively guanine rich includes the telomeres, and G4 DNA has been documented to form at those sequences (Patel et al., 2007). Finally, G4 DNA has been identified at oncogenes that

undergo frequent recombination and mutation. It has been shown that *c-MYC*, *HOX11* and *BCL2* all form G4 structures, implicating it in oncogenic translocations (Siddiqui-Jain et al., 2002; Nambiar et al., 2011, 2013).

While G4 DNA may play an important, although not fully defined role in regulating DNA and RNA activities, it also influences genome stability. This is because G4 is highly stable, and needs to be resolved so that it does not interfere with transcription or replication. Blooms syndrome patients, who are missing the Blooms Syndrome RecQ like (BLM) helicase are predisposed to cancer (Hickson et al., 2001), and BLM is a G4 helicase (Huber et al., 2002). A similar disease, Werner's syndrome, results in gross chromosomal rearrangements and a predisposition to cancer (Mohaghegh and Hickson, 2002), and is caused by defects in the WRN helicase, which also unwinds G4 DNA (Fry and Loeb, 1999). More recently, FANCJ helicase was shown to unwind G4 structures *in vitro* (Wu et al., 2008), and loss of that protein causes Fanconi's Anemia and defects in DNA repair (Wu and Brosh, 2010). Finally, Pif1 helicase is conserved throughout eukaryotes and suppresses the addition of *de novo* telomeres at double stranded breaks (Bochman et al., 2010).

A failure to resolve G4 DNA by helicases may promote instability, not only by creating a replication blockade, but also by inhibiting normal DNA repair processes. The mismatch repair (MMR) pathway is already known to process DNA repeats. MMR corrects replication errors, and its loss or disruption results in the instability of microsatellite repeats (Martin-Lopez and Fishel, 2013). Microsatellites are tandem repeats that are naturally polymorphic due to lowered

polymerase fidelity at repetitive DNA. Typically MMR suppresses those mutations, and loss of repair leads to observed site-specific increased instability at microsatellites. Indeed, microsatellite instability is a diagnostic tool for the loss of mismatch repair in certain heritable cancers such as non-polyposis colorectal cancer (Fishel et al., 1993, 1995). Extensive expansion of repeats past a certain threshold can also inactivate cancer genes, such as tumor suppressors (Markowitz et al., 1995), promoting cancer development (Souza et al., 1996).

While MMR loss is a well-known contributor to genome instability, there are other DNA repair pathways critical for genome maintenance. For example, nucleotide excision repair removes bulky DNA lesions such as UV crosslinks (Schärer, 2013), while homologous recombination and non-homologous end joining repair DNA breaks (Curtin, 2012). Another, base excision repair, corrects modified DNA bases, such as uracil (Krokan and Bjoras, 2013).

It is reasonable to question why genomes have retained unstable DNA sequences like those that support G4 DNA over evolutionary time. In short, it is not clear, but it does seem likely from available evidence that repetitive DNA strikes a balance between positive functional contributions to the cell and negative genome stability consequences. For example, DNA hairpin and G4 formation in prokaryotes and eukaryotes may be structural features at origins of replication (Kim and Sam, 1989; Lin and Kowalski, 1994; Cayrou et al., 2012; Wanarooij et al., 2010). These sequences are also found to be associated with gene regulatory domains in humans (Maizels, 2012; Bochman et al., 2012). Indeed, genome analyses showed that G4 forming regions are concentrated in

promoter regions, 5' UTRs and 5' introns, which are probably involved in gene regulation (Huppert and Balasubramanian, 2005, 2007; Zhao et al., 2010). The presence of guanine repeats in transcribed regions may help influence expression rate, benefits of which may outweigh any potential negative effect on genomic stability (Zhao et al., 2010). This interplay between formation and structure resolution could give cells the ability to regulate G4 structures and in turn regulate a multitude of cellular functions (Maizels, 2006, 2012; Siddiqui-Jain et al., 2002). This notion is backed by several studies showing G4 may be involved in a wide range of activities such as transcription, translation, recombination, replication initiation, aptamers (binding molecules), telomere maintenance, and mRNA processing (Larson et al., 2005; Marcel et al., 2011; Maizlels, 2006, 2012; Zybailov et al., 2013; Collie and Parkinson, 2011; Yoshida et al., 2013; Xu et al., 2010; Blackburn, 1991). Regulation is likely due to the structure and not simply the sequence because site-directed mutagenesis targeted to disrupt G4 formation altered the regulatory capacity (Marcel et al., 2011; Siddiqui-Jain et al., 2002; Nambiar et al., 2013).

Given that guanine-rich sequences support the formation of G4 DNA, and that G4 DNA promotes instability, it is logical to predict that genome sequences containing G4 motifs may be prone to recombination or mutagenesis. While this rationale follows with contemporary research, clear connections between specific regions of instability and G4 DNA have yet to be drawn. Therefore, my dissertation will examine the connection between G4 DNA and site-specific instability. Using computational approaches in Chapter 2, I focus on identification

and investigation into new biologically relevant G4 loci in the human genome. A subset of novel large G4 loci was assayed for *in vitro* G4 formation using multiple complimentary techniques. Using human sequence variation databases, these loci showed signs of increased mutagenesis on both a small and large-scale. At the experimental level, in Chapter 3 I zeroed in on one particular gene and examined the sources of its instability. This gene is called *TCF3*, and it is a major regulatory protein that participates in oncogenic translocations. The molecular mechanism of *TCF3* instability is unknown, and my data suggest that G4 DNA is involved. Chapter 4 addresses molecular mechanisms of G4 instability. I examined how mismatch repair factors respond to the presence of G4 DNA structures. G4 sequences in the human genome may be hot spots for mutagenesis, and this may be due in part to a failure of normal repair. Together, this dissertation provides molecular insights into the apparent genetic instability at repetitive guanine rich sequences and provides a new perspective on the biological impact of structure formation.

References

Abeysinghe, S. S., Chuzhanova, N., Krawczak, M., Ball, E. V., & Cooper, D. N. (2003). Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination‐associated motifs. *Human mutation*, 22(3), 229-244.

Balasubramanian S., and Neidle S. (2009). G-quadruplex nucleic acids as therapeutic targets. *Current opinion in chemical biology*, *13*(3), 345-353.

Bacolla, A., Wojciechowska, M., Kosmider, B., Larson, J. E., & Wells, R. D. (2006). The involvement of non-B DNA structures in gross chromosomal rearrangements. *DNA repair*, 5(9), 1161-1170.

Biffi, G., Tannahill, D., McCafferty, J., & Balasubramanian, S. (2013). Quantitative visualization of DNA G-quadruplex structures in human cells. *Nature Chemistry*, *5*(3), 182-186.

Blackburn, E. H. (1991) Structure and function of telomeres. *Nature* 350:569-573.

Bochman, M. L., Sabouri, N., & Zakian, V. A. (2010). Unwinding the functions of the Pif1 family helicases. *DNA repair*, *9*(3), 237-249.

Bochman, M.L., Paeschke, K., Zakian V.A. (2012). DNA secondary structures: stability and function of G-quadruplex structures. *Nature Reviews Genetics*. 13:770-780.

Brooks, T.A., Kendrick, S., Hurley, L. (2010). Making sense of G-quadruplex and i- motif functions in oncogene promoters. *Febs* Journal 277:3459-3469.

Bunting, S. F., & Nussenzweig, A. (2013). End-joining, translocations and cancer. *Nature Reviews Cancer*, *13*(7), 443-454.

Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K., & Neidle, S. (2006). Quadruplex DNA: sequence, topology and structure. *Nucleic acids research*, 34(19), 5402-5415.

Cayrou, C., Coulombe, P., Puy, A., Rialle, S., Kaplan, N., Segal, E., & Méchali, M. (2012). New insights into replication origin characteristics in metazoans. *Cell Cycle,* 11(4), 658-667.

Cogoi, S., Paramasivam, M., Spolaore, B., & Xodo, L. E. (2008). Structural polymorphism within a regulatory element of the human KRAS promoter: formation of G4-DNA recognized by nuclear proteins. *Nucleic Acids Research*, 36(11), 3765-3780.

Collie, G., Parkinson, G.N. (2011). The application of DNA and RNA G-quadruplexes to therapeutic medicines. *Chemical Society Reviews* 40: 5867-5892.

Curtin, N.J., (2012). DNA repair dysregulation from cancer driver to therapeutic target. *Nature Reviews Cancer* 12:801-817.

Duquette, M. L., Handa, P., Vincent, J. A., Taylor, A. F., & Maizels, N. (2004). Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. *Genes & development*, 18(13), 1618-1629.

Federici, L., Arcovito, A., Scaglione, G. L., Scaloni, F., Sterzo, C. L., Di Matteo, A., Brunori, M. (2010). Nucleophosmin C-terminal leukemia-associated domain interacts with G-rich quadruplex forming DNA. Journal of *Biological Chemistry*, 285(48), 37138-37149.

Fishel, R., & Kolodner, R. D. (1995). Identification of mismatch repair genes and their role in the development of cancer. *Current opinion in genetics & development*, 5(3), 382-395.

Fishel, R., Lescoe, M. K., Rao, M. R. S., Copeland, N. G., Jenkins, N. A., Garber, J., Kolodner, R. (1993). The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell*, *75*(5), 1027-1038.

Fry, M., & Loeb, L. A. (1999). Human werner syndrome DNA helicase unwinds tetrahelical structures of the fragile X syndrome repeat sequence d (CGG) n. J*ournal of Biological Chemist*ry, 274(18), 12797-12802.

Gellert, M., Lipsett, M. N., & Davies, D. R. (1962). Helix formation by guanylic acid. *PNAS*, *48*(12), 2013.

González, V., Guo, K., Hurley, L., & Sun, D. (2009). Identification and characterization of nucleolin as a c-myc G-quadruplex-binding protein. *Journal of Biological Chemistry*, 284(35), 23622-23635.

Gray, D. M., & Bollum, F. J. (1974). A circular dichroism study of poly dG, poly dC, and poly dG: dC. *Biopolymers*, *13*(10), 2087-2102.

Hickson, I. D., Davies, S. L., Li, J. L., Levitt, N. C., Mohaghegh, P., North, P. S., & Wu, L. (2001). Role of the Bloom's syndrome helicase in maintenance of genome stability. *Biochemical Society Transactions*, *29*(2), 201-204.

Huber, M. D., Lee, D. C., and Maizels, N. (2002). G4 DNA unwinding by BLM and Sgs1p: substrate specificity and substrate-specific inhibition. *Nucleic Acids Research*, *30*(18), 3954-3961.

Huppert, J. L., & Balasubramanian, S. (2005). Prevalence of quadruplexes in the human genome. *Nucleic acids research*, *33*(9), 2908-2916.

Huppert, J. L., & Balasubramanian, S. (2007). G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res*earch 35:406-413.

Katapadi, V. K., Nambiar, M., & Raghavan, S. C. (2012). Potential G-quadruplex formation at breakpoint regions of chromosomal translocations in cancer may explain their fragility. *Genomics*, 100(2), 72-80.

Kennedy, S. R., Loeb, L. A., & Herr, A. J. (2012). Somatic mutations in aging, cancer and neurodegeneration. *Mechanisms of ageing and development*, *133*(4), 118-126.

Kim, Y. S., & Kang, H. S. (1989). Sequence-specific functions of the early palindrome domain within the SV40 core origin of replication. *Nucleic Acids Res*earch 17: 9279-9289.

Koole, W., van Schendel, R., Karambelas, A. E., van Heteren, J. T., Okihara, K. L., & Tijsterman, M. (2014). A Polymerase Theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites. *Nature* communications 5.

Krokan, H. E., & Bjørås, M. (2013). Base excision repair. *Cold Spring Harbor perspectives in biology*, 5(4), a012583.

Larson, E. D., Duquette, M. L., Cummings, W. J., Streiff, R. J., & Maizels, N. (2005). MutSα binds to and promotes synapsis of transcriptionally activated immunoglobulin switch regions. *Current biology*, *15*(5), 470-474.

Lin, S., & Kowalski, D. (1994). DNA helical instability facilitates initiation at the SV40 replication origin. *Journal of molecular biology* 235:496-507.

Loeb, L. A. (2011). Human cancers express mutator phenotypes: origin, consequences and targeting. *Nature* Reviews Cancer, 11(6), 450-457.

Maizels, N., (2006). Dynamic roles for G4 DNA in the biology of eukaryotic cells. *Nature structural & molecular biology* 13:1055-1059.

Maizels, N., (2012). G4 motifs in human genes. *Annals of the New York Academy of Sciences* 1267: 53-60.

Marcel, V., Tran, P. L., Sagne, C., Martel-Planche, G., Vaslin, L., Teulade-Fichou, M. P., Van Dyck, E. (2011). G-quadruplex structures in TP53 intron 3: role in alternative splicing and in production of p53 mRNA isoforms. *Carcinogenesi*s, 32(3), 271-278.

Markowitz, S., Wang, J., Myeroff, L., Parsons, R., Sun, L., Lutterbaugh, J., Vogelstein, B. (1995). Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. *Science*, *268*(5215), 1336-1338.

Martín-López, J. V., & Fishel, R. (2013). The mechanism of mismatch repair and the functional analysis of mismatch repair defects in Lynch syndrome. *Familial cancer*, *12*(2), 159-168.

McMurray, C.T., (2010). Mechanisms of trinucleotide repeat instability during human development. *Nature Reviews Genetics* 11:786-799.

Mohaghegh, P., & Hickson, I. D. (2002). Premature aging in RecQ helicase-deficient human syndromes. *The international journal of biochemistry & cell biology*, *34*(11), 1496-1501.

Nambiar, M., Srivastava, M., Gopalakrishnan, V., Sankaran, S. K., & Raghavan, S. C. (2013). G-quadruplex structures formed at the HOX11 breakpoint region contribute to its fragility during t (10; 14) translocation in T-cell leukemia. *Molecular and cellular biology*, 33(21), 4266-4281.

Nag, D.K., Petes, T.D. (1991). Seven-base-pair inverted repeats in DNA form stable hairpins *in vivo* in Saccharomyces cerevisiae. *Genetics* 129: 669-673.

Negrini, S., Gorgoulis, V. G., & Halazonetis, T. D. (2010). Genomic instability—an evolving hallmark of cancer. *Nature reviews Molecular cell biology*, 11(3), 220-228.

Nguyen, G. H., Tang, W., Robles, A. I., Beyer, R. P., Gray, L. T., Welsh, J. A., ... & Harris, C. C. (2014). Regulation of gene expression by the BLM helicase correlates with the presence of G-quadruplex DNA motifs. *PNAS* 111:9905-9910.

Patel, D. J., Phan, A. T., & Kuryavyi, V. (2007). Human telomere, oncogenic promoter and 5′-UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics. *Nucleic Acids Res*earch, 35(22), 7429-7455.

Piazza, A., Boulé, J. B., Lopes, J., Mingo, K., Largy, E., Teulade-Fichou, M. P., and Nicolas, A. (2010). Genetic instability triggered by G-quadruplex interacting Phen-DC compounds in Saccharomyces cerevisiae. *Nucleic Acids Research*, *38*(13), 4337-4348.

Piazza, A., Serero, A., Boule, J. B., Legoix-Ne, P., Lopes, J., and Nicolas, A. (2012). Stimulation of gross chromosomal rearrangements by the human CEB1 and CEB25 minisatellites in Saccharomyces cerevisiae depends on G-quadruplexes or Cdc13. *PLoS genetics*, *8*(11), e1003033.

Pikor, L., Thu, K., Vucic, E., & Lam, W. (2013). The detection and implication of genome instability in cancer. *Cancer and Metastasis Reviews*, *32*(3-4), 341-352.

Ralph, R. K., Connors, W. J., & Khorana, H. G. (1962). Secondary structure and aggregation in deoxyguanosine oligonucleotides. *Journal of the American Chemical Society*, 84(11), 2265-2266.

Renton, A. E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J. R., Sulkava, R. (2011). A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. N*euron*, 72(2), 257-268

Ribeyre, C., Lopes, J., Boulé, J. B., Piazza, A., Guédin, A., Zakian, V. A., ... & Nicolas, A. (2009). The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. PLoS genetics, 5(5), e1000475.

Schärer, O. D. (2013). Nucleotide excision repair in eukaryotes. *Cold Spring Harbor perspectives in biology*, *5*(10), a012609.

Sen, D., Gilbert, W. (1988). Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature* 364-366.

Sen, D., Gilbert, W. (1990). A sodium-potassium switch in the formation of four-stranded G4-DNA. *Nature* 344:410-414.

Sen, D., Gilbert, W. (1992). Novel DNA superstructures formed by telomere-like oligomers. Biochemistry 31:65-70.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research 29*:308-311.

Siddiqui-Jain, A., Grand, C. L., Bearss, D. J., & Hurley, L. H. (2002). Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *PNAS*, 99:11593-11598.

Souza, R. F., Appel, R., Yin, J., Wang, S., Smolinski, K. N., Abraham, J. M., Meltzer, S. J. (1996). Microsatellite instability in the insulin–like growth factor II receptor gene in gastrointestinal tumours. *Nature* genetics 14:255-257.

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S.,Cooper, D. N. (2003). Human gene mutation database (HGMD®): 2003 update. *Human mutation*, 21(6), 577-581.

Sun, H., Karow, J. K., Hickson, I. D., & Maizels, N. (1998). The Bloom's syndrome helicase unwinds G4 DNA. Journal of Biological Chemistry, 273(42), 27587-27592.

Tarsounas, M., Tijsterman, M. (2013). Genomes and G-quadruplexes: for better or for worse. *Journal of molecular biology* 425:4782-4789.

Todd, A.K., Johnston, M., Neidle, S. (2005). Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res* 33:2901–2907.

van Kregten, M., Tijsterman, M. (2014). The repair of G-quadruplex-induced DNA damage. *Experimental cell research* 329:178-183.

Wanrooij, P. H., Uhler, J. P., Simonsson, T., Falkenberg, M., & Gustafsson, C. M. (2010). G-quadruplex structures in RNA stimulate mitochondrial transcription termination and primer formation. *PNAS* 107(37), 16072-16077.

Wells, R. D. (2007). Non-B DNA conformations, mutagenesis and disease. *Trends in biochemical sciences* 32:271-278.

Wu, Y., Shin-ya, K., Brosh, R.M. (2008). FANCJ helicase defective in Fanconia anemia and breast cancer unwinds G-quadruplex DNA to defend genomic stability. *Molecular and cellular biology 28:*4116-4128.

Wu, Y., Brosh, R.M. (2010). G-quadruplex nucleic acids and human disease. *FEBS journal* 277:3470-3488.

Xu, L., Zhang, D., Huang, J., Deng, M., Zhang, M., Zhou, X. (2010). High fluorescence selectivity and visual detection of G-quadruplex structures by a novel dinuclear ruthenium complex. Chemical Communications 46:743-745.

Yadav, P., Harcy, V., Argueso, J. L., Dominska, M., Jinks-Robertson, S., & Kim, N. (2014). Topoisomerase I Plays a Critical Role in Suppressing Genome Instability at a Highly Transcribed G-Quadruplex-Forming Sequence. *PLoS genetics* 10:e1004839.

Yoshida, W., Saito, T., Yokoyama, T., Ferri, S., & Ikebukuro, K. (2013). Aptamer selection based on g4-forming promoter region. *PloS one* 8: e65497.

Zhao, J., Bacolla, A., Wang, G., & Vasquez, K. M. (2010). Non-B DNA structure-induced genetic instability and evolution. *Cellular and molecular life sciences*, 67(1), 43-62.

Zybailov, B. L., Sherpa, M. D., Glazko, G. V., Raney, K. D., & Glazko, V. I. (2013). G4-quadruplexes and genome instability. *Molecular Biology*, 47(2), 197-204.

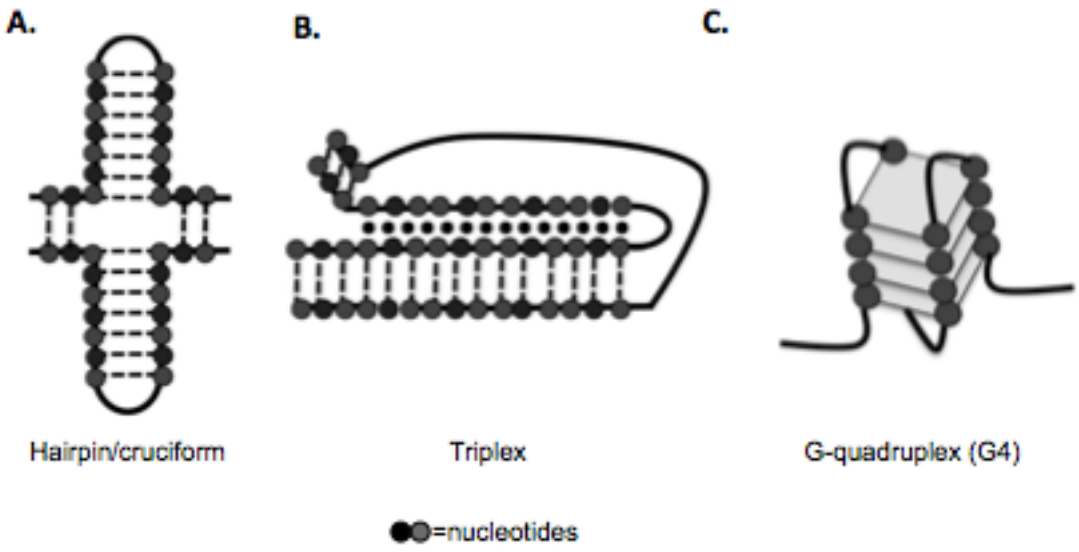

Figure 1. Subset of Non-B Form DNA Secondary Structures Capable of Forming in the Human Genome. Dots indicate nucleotides involved in structure formation. Dotted lines, smaller dots, and gray boxes represent different base pair interactions responsible for forming secondary structures. Solid lines are nucleotides not involved in structure formation. **(A)** Cruciform structures form from hairpin formations on both strands. **(B)** Triplex structure formation is a three stranded fold back structure with alternative base pair interactions (•) are specific to high pH conditions. **(C)** G-quadruplex (G4) is a four-stranded structure that uses alternative guanine hydrogen bonds for stability (gray box).

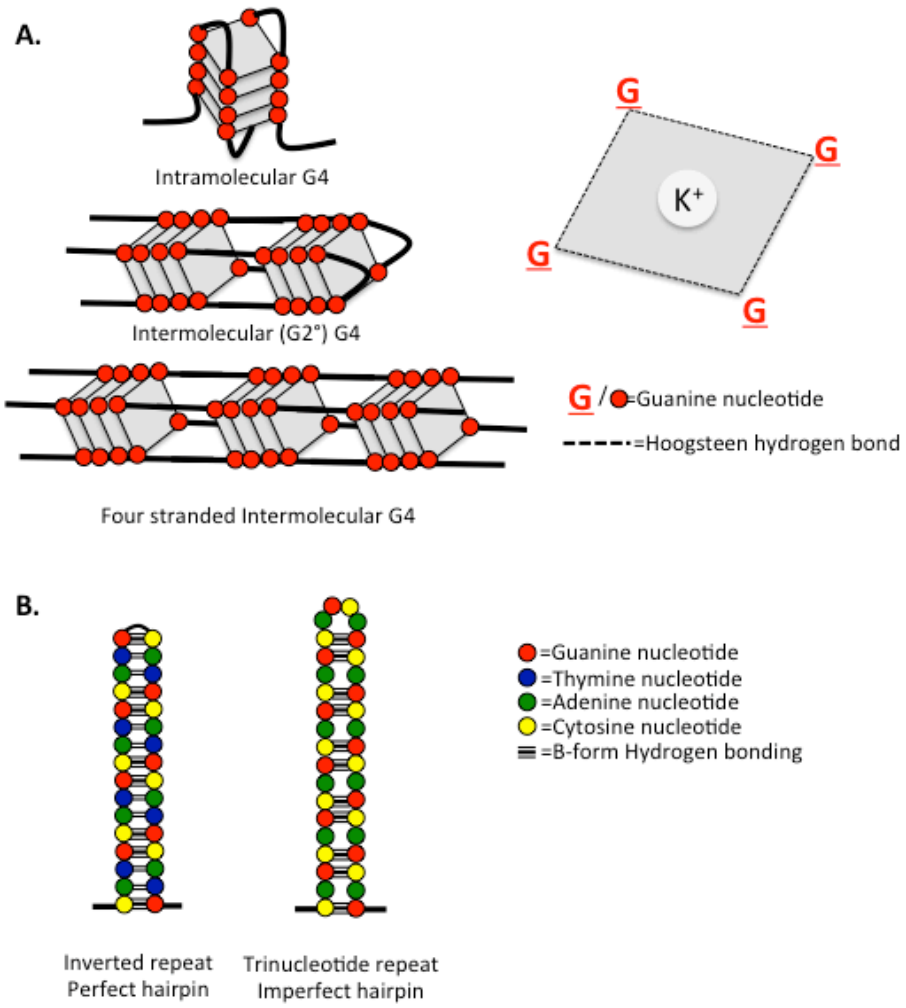

Figure 2. Different Sequence Compositions Allow Vastly Different DNA Structure Formations. Specific nucleotide compositions allow a dynamic array of structures to form. Each type of nucleotide is represented by a specific color dot with guanine = red, thymine = blue, adenine = green, and cytosine=yellow. **(A)** Depiction of the different types of G4 structures able to form in the human genome. G4 is formed by four guanine (G / red dot) Hoogsteen hydrogen bond interactions (dotted line) that form a tetrad (right), and multiple tetrads are stacked on top of each from consecutive G-repeats (left). G4 can form intra-molecular (single-strand) or inter-molecular (multiple strands) structures depending on the sequence composition and cellular conditions. **(B)** Depiction of the different types of hairpins capable of forming in the human genome. Inverted repeats allow alternative B-form hydrogen bonding between every base pair (left, perfect hairpin). Trinucleotide repeats (shown CAG) form imperfect hairpins with hydrogen bonds between C-G nucleotides only.

CHAPTER II

IDENTIFICATION AND CHARACTERIZATION OF LARGE G-QUADRUPLEX SEQUENCES IN THE HUMAN GENOME

Abstract

Previous computational analysis of the human genome has predicted over 300,000 motifs that are capable of forming G4 structures. Many of these regions are involved in regulation and potentially site-specific mutagenesis. Longer repeats are generally more stable. The extent and impact of such sequences on human genome instability is not well characterized. In this chapter, I present results form bioinformatic analyses that identify and characterize a panel of the largest G4 motifs present in the human genome, which I call large G4 (LG4s). I found that LG4 regions are prone to sequence variations on both the large and small-scale. My results add support to a hypothesis that sequences containing G4 repeats promote mutagenesis. This study connects large G4 motifs with genomic changes. It also identifies new genetic candidates for instability caused by G4 structure formation. These results have profound implications for understanding the molecular basis of multiple unrelated human diseases because it connects structure formation with site-specific instability.

Introduction

A majority of the human genome (~98%) is composed of repetitive or non-coding DNA (Elgar and Vavouri, 2008). However, the distribution of repeats is not random, and there is evidence suggesting that repetitive sequences trigger DNA rearrangements and mutagenesis. This is probably attributable to an ability for DNA repeats to support non-B form DNA conformations. In particular, guanine repeats can fold into four-stranded G-quadruplex (G4) DNA. In this chapter I will report on my bioinformatics analysis of frequent site-specific rearrangements in the human genome and their connections with G4 DNA.

The ability of a given sequence to form G4 DNA can be predicted based on the sequence alone. By applying known parameters for G4 formation, a number of structure predication programs have been developed. Taking into account loop length and the number guanine repeats needed as a minimal G4 sequence, it has been estimated that yeast genomes hold over 1,400 individual G4 motifs (Capra et al., 2010), whereas over 300,000 motifs potentially exist in the human genome (Huppert and Balasubramanian, 2005; Todd et al., 2005). G4 DNA has even been detected in prokaryotic genomes too, with current studies suggesting a range of 45 to 30,000 individual G4 motifs, depending on the species (Rawal et al., 2006). Interestingly, many G4 motifs are evolutionarily conserved within human populations as well as between some related yeast species (Capra et al., 2010; Nakken et al., 2009), so G4 DNA may impart some functional benefit to a cell.

Informatic research on unstable loci suggests that G4 DNA promotes instability. A computational analysis of all known translocation breakpoints indicates that 70% occur at sequences capable of supporting G4 DNA (Katapadi et al., 2012). Analyses of small sequence variation databases indicate that G4 sequences could be hot spots for mutagenesis (Du et al., 2014). Next generation sequencing studies of individuals with amyotrophic later sclerosis (ALS) resulted in the identification of an intronic G-rich repeat whose length was expanded over 100 times compared to healthy individuals (Renton et al., 2011). This G-rich repeat was recently shown to form G4 in RNA as well as DNA, and is now implicated in the etiology of sporadic and familial ALS (Haeusler et al., 2014).

While informatics has started to characterize the location and abundance of G4 loci in the human genome, connections between guanine-rich DNA and site-specific mutagenesis are only beginning to be recognized. In this chapter, I present results from my analysis of a panel of the largest G4 motifs present in the human genome, which I call large G4 (LG4s). Reported here are findings showing that LG4 regions appear to be prone to sequence variations on both large and small-scales, supporting earlier models regarding G4 DNA and instability. I also provide evidence that G4 repeat expansions may occur in human populations. These results have profound implications for understanding the molecular basis of multiple, unrelated, human diseases containing a genome instability component.

Materials and Methods

Java-Based Program for Identification of LG4

In order to search and identify a dataset of the largest G4 regions (LG4s) present in the human genome, we wrote a Java program to search for any FASTA sequence file for G-triplet regions likely to form G4 (Unpublished program, designed by Brad Dyninski and Glen Borchert). Our program is unique because does not take in account loop length, but does identify G4 motifs based on the density of G-triplets within 1.5 (kb) sequence windows. These windows overlap to avoid missing sequences that would otherwise transverse two sequence blocks. In order to establish the minimal density of G-triplets needed for G4 formation, we designed a search program based on a long model G4 sequence found at the human immunoglobulin switch region (Sγ3) (Ehrat et al., 2012; Duquette et al., 2004). Overlapping sliding windows were applied, with a minimal output threshold of (GGGn) X 120 for every 1.5 kilobase (kb) of sequence.

Sequence Analyses

Output of the Java LG4 identification program was used to map each individual LG4 location on Ensembl Release 69 (hg19) and Release 77 (ch38) (Flicek et al., 2011). The LG4 locations on corresponding diagrams were obtained by using Ensembl77 (Kersey et al., 2014). Potential regulatory functions and ChipSeq pull down data was obtained using Ensembl77 with the Regulatory Build filter turned on (Kersey et al., 2014).

A program called QGRS mapper (Kikin et al., 2006) was used to determine the potential of each sequence to form G4. This was accomplished using the following filters: A max length loop length of 45 nucleotides, min G group of 3, and a loop size 0-36 nucleotides (selects for intra-molecular G4 only). The output of the analysis was mapped to the location of the LG4, and the number of individual non-overlapping G4 motifs for every kilobase (kb) (G4 density) was calculated for each LG4. Additionally, positions directly adjacent of LG4 at 1 kb increments, 5' and 3' up to 5 kb away, and the remaining transcript over 5 kb away from LG4s were also calculated.

I designed a second program using Perl programming language to calculate the composition of guanine repeats, shown in Figure 5. Briefly, LG4 repeats were formatted into a FASTA file containing only the repetitive G-rich sequences. Only LG4s found in protein transcript regions were used in this analysis. The data obtained from the program was used to calculate each LG4's repeat composition of the 3 most prominent G- repeats potentially involved in G4 structure formation; 3, 4 or 5 guanine mononucleotide repeats.

Copy number variations (CNVs) for LG4s found in protein transcript regions were downloaded from the database of genomic structural variation (dbVAR) on NCBI.org (Lappalainen et al., 2013). The density of CNV breakpoints (CNV > 99 base pairs) was calculated using the exact reference points as the density of G4 motifs above.

The location of all individual insertions and deletions were obtained from the dbSNP database (Sherry et at., 2001) and mapped to LG4s and surrounding

introns using Enseml release 69 (Flicek et al., 2011). The density of small sequence variants was calculated by number of insertion or deletion events per 100 base pairs (bp) for LG4 and regions 1-2000 bp away from LG4s.

Statistical Analysis

All statistical analyses were performed using StatPlus:macV5. If the analysis contained more than 2 variables, a one-way Anova was used to calculate statistical significance, unless specified otherwise. Comparisons of two variables used an unpaired two-tailed *t*-test. All *p* values from statistical analysis are shown on the corresponding graph.

Circular Dichroism

Oligonucleotides for Circular Dichroism (CD) studies were designed by using representative repeat units found in LG4 sequences, and were synthesized by Operon (Eurofins MWG operon LLC, Huntsville, AL 35805). CD analysis was performed using an Aviv model 215 CD spectrometer at 37°C. Spectra were taken in 1 cm path quartz cells containing 12 μM G4 oligonucleotide in 10 mM Tris–HCl, pH 7.6, 1 mM EDTA, and 100 mM KCl. The molar ellipticity was measured from 220–300 nm and recorded for 3 scans in 1 nm increments at a 1 sec average time.

Primer Extension Assays

LG4 containing phagemids for extension assays were obtained by cloning PCR amplified genomic fragments, or cloned from amplification products using overlapping primers in a standard PCR reaction. PCR products were gel purified and TOPO cloned (Invitrogen, Life Technologies, Carlsbad, CA) into pCR2.1. Fragments were cloned in both orientations and were verified by Sanger sequencing (University of Illinois Core Sequencing Center). Templates for extension assays are shown in Table 8 and include additional genome sequence surrounding G4 repeat, ranging from 120-600 base pairs (bp). Closed-circular single-stranded DNA was obtained using M13K07 helper phage (NEB) according to the manufacture's instructions.

Klenow Polymerase extension assays were performed essentially as described in (Ehrat et al., 2012) and based on previous G4 assays (Sun and Hurley, 2010; Weitzmann et al., 1996). Single-stranded phagemid templates were primed with a ^{32}P 5' end labeled M13 forward primer, which was then extended with Klenow (NEB). In addition to the manufacture's buffer, KCl or LiCl was added to a final concentration of 25 mM. Klenow extension reactions were performed at 37 °C for 8 minutes with 5' end-labeled M13 forward (-20) primer. Extension reactions were stopped by the addition of an equal volume of 90% formamide and 1 mM EDTA followed by heating to 90°C for 20 minutes. Products of polymerase extension were resolved by 8% denaturing PAGE (19:1) with 7 M urea and 0.5X TBE, at 700 V at room temperature. Gels were then dried and

images were captured by phosphorimaging using a Molecular Dynamics Storm 840 phosphorimager (Amersham/GE).

PCR of *CRLF2* LG4 from Human Genomic Template

PCR was preformed using human disease free genomic DNA (Invitrogen) using Taq Polymerase (NEB) according to the manufacturer's instructions. PCR products were separated on a 1% agarose gel and 1kb DNA ladder (NEB) was used to identify product sizes. The 1.2 kb band was gel excised, TOPO Cloned (Invitrogen), and verified by sequencing.

Identification of Disease Genes

All proteins containing an LG4 sequence in their genomic DNA were analyzed on the Database for Annotation, Visualization and Integrated Discovery (DAVID), a web based program that provides annotation tools for researchers to understand the biological meaning behind large list of genes identified in microarray or bioinformatic studies (Huang et al., 2008). DAVID can be used to identify protein interactions, common pathways, disease relevance, and multiple other analyses. Wiki-gene web interface (Wu et al.,2009) was also used to search primary literature for LG4-proteins involved in disease not listed on DAVID's database.

Results

Identification of Large G4 Regions in the Human Genome

Most G4 prediction programs use an algorithm based on a minimal definition for the formation of G4, which results in identification of an unmanageable number of sequences for a given genome. We set out to identify a panel of loci containing extensive G4 sequence motifs, thereby focusing on regions of the genome that are highly repetitive in guanine. Long guanine-rich minisatellites (Piazza et al., 2010) and the 2-10 kb guanine-rich switch regions (Maizels, 2006) both adopt G4 DNA, and are associated with DNA breaks. Therefore, we reasoned that guanine repeats covering a sequence of at least 1 kb would be particularly unstable and that these should not be overly abundant in the genome. A Java script program was written to scan genomic regions and count the number of G-triplets. The size of this region scanned is determined by the size of the "sliding window", which can be readily changed in the program. From here we selected a known G4 region, the Sγ3 intron, as a model for "large G4" sequences and based our program on its guanine density. The output of the program is genomic regions that contain a high density of G-triplets (120 G-triplets/1.5kb window). This program was applied to the entire human genome, and the G4 loci we found are described in this chapter. The "LG4" term is used here as a name given to G4 sequences present in the human genome containing >120 GGG repeats/1500 base pairs.

We were able to identify 315 large G4 capable regions (LG4s) in the human genome. The sizes varied widely, ranging from ~600–7,000 nucleotides

long with diverse genomic locations. A majority of LG4s were found in transcriptional regulatory locations, with 49% of LG4s within protein transcripts, and 8% within 2 kb of a transcribed protein-coding region (Table 1). There was a slight trend for G-rich sequences to be on the transcribed strand (CCC mRNA) over non-transcribed strand (Table 3), with 92 compared to 62 loci, respectively.

Previous reports identified G4 sequence motifs enriched at promoter and 5' intronic regions (Eddy and Maizels 2008; Huppert and Balasubramanian 2007). Our search returned similar results with over half (54%) of LG4s located in promoter or 5' intronic regions (UTR+ 5' intron+ 2 kb 5' of UTR). Surprisingly, 25% were found in 3' and internal introns (Table 1-2). While promoter regions and 5' G4 introns have been suggested to help regulate transcription and translation, little is known about the role of G4 DNA found in 3' regions; although some studies have suggested roles in mRNA localization, mRNA splicing, and miRNA interference (Beaudoin et al., 2013; Subramanian et al., 2011; Arora and Suess, 2011). Further characterization of the 3' motifs could provide new insights into the regulatory capability of G4.

Validation of G4 Folding Potential

Quadruplex forming G-Rich Sequences (QGRS) mapper is a web-based program for identifying individual G4 motifs in a given DNA sequence (Kikin et al., 2006). Full protein transcripts containing LG4 sequences were queried with QGRS in both orientations; then the average number of non-overlapping G4 motifs/1 kb (G4 motif density) was calculated. This was completed for individual

LG4s, and 5 kb on either the 5' and 3' sides of the LG4 sequence (at 1 kb increments). I also analyzed the transcript not associated with LG4. On average, LG4s contained 18 individual G4 motifs/kb (red line at LG4 Figure 4A), a significant (9-fold) increase (p=1.3X10^{-113}) compared to loci directly flanking LG4 (5'-1 and 3'-1 Figure 4A) and transcripts not in proximity to LG4s (dotted green line Figure 4A). Such a high average density and significance shows that the regions I analyzed contained multiple G4 sequence motifs.

Due to a sufficient sample size in each mRNA location, I aimed to test if location of the LG4 in the transcript had any correlation with the relative density of G4 motifs. G4 motif density is defined here as the average number of non-overlapping G4 motifs for every 1kb of DNA sequence. The locations of LG4 in the mRNA were grouped into the following categories: If located in the UTR (UTR-LG4), in an UTR at the 5' end of the gene (5' UTR), in the first two 5' introns (5'-LG4), internal intron (Middle-LG4), or the 3' intron (3'-LG4,). 3' UTR-LG4s were removed from analysis due to small sample size. These are the groupings (and nomenclature) I will refer to throughout this chapter and dissertation.

I found no relationship between individual G4 motif density in LG4 and its location in the mRNA (LG4 central x-axis Figure 4B). However, there was a significant increase (p=.006) in the density of G4 motifs located directly 3' of 5'UTR-LG4s (5' UTR blue line, 1-3' Figure 4B). This is in agreement with previous analysis that 5' intronic regions contain G4 capable sequences (Eddy and Maizels 2008). My results also indicate that the position of a G4 sequence

within an mRNA occurs independent of the density of G4 motifs, or, the density of the G4 motifs has no relationship to where the repeat is positioned in the gene.

LG4s are found in Sequences Involved in Gene Regulation

Ensembl.org is a multifaceted web interface that allows access and visualization of databases with respect to their location in the human genome. Ensembl Regulatory Build Database is a collection of regulatory regions in the human genome that are categorized by gene regulatory elements, such as predicted promoter regions, promoter flanking regions, predicted enhancers, CTCF binding sites, transcription factor binding sites, and open chromatin regions (OC) (Flicek et al., 2013). The database is composed of regions that were identified using the publicly available experimental data sets from DNase1-Seq, FAIRE-Seq, and ChIP-Seq studies on the human genome, and then uploaded to Ensembl77.

I investigated LG4 sequences for transcribed regulatory elements, and found that over half are correlated with regions associated with gene regulation (Table 4). Promoter-associated regions and open chromatin states were the most prominent LG4 regulatory elements, with 36 and 29 respectively (Table 5). Consistent with previous studies, LG4 sequences localized within promoters were found primarily in the UTR and 5' introns (Table 5) (Huppert and Balasubramanian, 2007; Eddy and Maizels, 2008). To my surprise, LG4 was also located in open chromatin regions (OC) (15/29), particularly in middle introns (Table 5). OC regions could permit regulatory interactions (Kumar et al., 2009) or be sites for increased instability (Folle, 2008). Only one gene, Tetra-Peptide

Repeat Homeobox 1 (*TPRX1*), was found to contain a LG4 sequence in an exon (Table 5).

Through analysis of Regulatory Build Chip-Seq data, 36 LG4 sequences showed evidence of interactions (72) with 20 different types of transcription factors (Table 6). Previous reports showed that the Specificity Protein 1 (*SP1*) transcription factor interacts with G4 promoter regions (Eddy and Maizels, 2008), and all 7 SP1 LG4 interacting motifs found were in promoter regions (Table 7). The dominant LG4-interacting transcription factor was Early Growth Response Protein 1 (*EGR1*) (Table 6). Similar to *SP1, EGR1* interactions were primarily in UTR and 5' promoters (Table 7.1). Multiple LG4-*EGR1* interactions were also observed in middle and 3' promoters, as well as OC regions, suggesting involvement in transcription of 3' G4 sequences (Table 7). These findings support previous research that G4 DNA is involved in gene regulation. It also identifies *EGR1* as a transcription factor that may interact with large G4 sequences.

The importance of LG4s in regulation is best exemplified by a close investigation of Max Interactor 1 (*MXI1*), a MYC family protein frequently mutated in prostate cancer (Eagle et al., 1995; Taj et al., 2001). Using Ensemlbe77 (Flicek et al., 2013), I found that the *MXI1*'*s* 5' LG4 intron is in a promoter region, and has *Sp1* and *Egr1* transcription factor interactions. Astonishingly, 4/6 of the alternative protein coding transcripts initiate directly 5' or 3' of the LG4 intron (Figure 3). Therefore, it is possible that *SP1* and *EGR1* transcription factor interactions at *MXI1*'s LG4 region influences isoform production.

G4 DNA in Exons

TPRX1 emerged as the only gene containing LG4 in an exon. *TPRX1* is a primate specific (Ensembl77) homeobox gene associated with development (Booth and Holland, 2007), and is part of a larger family of key regulatory proteins (Samuel et al., 2005). Although the mRNA is C-rich, 48% of the coding region is composed of the LG4 motif (1000/2076 base pairs). It is possible that *TPRX1*'s LG4 structure may be involved in regulating the proteins expression. Further analysis on Ensembl Regulatory Build supported a potential regulatory role. A dnase1-seq assay detected this region in an open chromatin state in H1ESC stem cells, and was possibly a transcription factor-binding site (Kersey et al., 2014).

The mRNA Location and Regulatory Function Influences LG4 Guanine Repeat Compositions

With the LG4 regulatory potential studied, I next asked if the guanine composition of G4 regions could account for their diverse regulatory presence in the human genome, and further, if this had any effect on site-specific genetic instability. The composition of G4 motifs in a given G-rich sequence can be complex, and little is known about the impact of different sequences on instability, or if any exist. In order to address this deficiency, a Perl-based program was designed to calculate individual LG4 guanine compositions of 3, 4, and 5-mononucleotide guanine repeats (Figure 5). The reason 3, 4, and 5-mononucleotide repeats were selected is because they are the predominant repeat required for G4 structure

formation. The program also counted the number of individual guanines regardless of whether it was located in a mononucleotide G-repeat or not, and is referred to here as the percent guanine (%G). I was particularly interested in deciphering the complexities of guanine repeat compositions within G4 to determine if patterns emerged among large G4 loci in the human genome. The density of G4 motifs/kb was compared to the percentage of total guanine base pairs (%G) (Figure 5A) to determine if the amount of guanine composing the repetitive sequence reflected G4 density of expressed LG4s. A significant relationship (p=1.4X10^{-6}) was found between the %G and individual G4 motif density (Figure 6A). Similar to G4 density (Figure 4B, represented as a bar graph Figure 6C), the percentage of guanines had no influence on LG4s location within the mRNA (Figure 6B).

Guanine repeats that support G4 structure formation generally contain 3, 4, or 5 tandem guanines, with more stable G4 structures containing much longer G-repeats (Maizels, 2006; Burge et al., 2006). To calculate the composition of each guanine tract for LG4s, mononucleotide G-repeats of 3, 4, or 5 were changed into countable characters (ex. X, Y, Z) using Perl programming language (Figure 5B). The program output displays the repeat composition as a percentage into Microsoft Excel for calculation (# of specific G-repeats X # of G's in a repeat / length of LG4). A strong statistical relationship was found between the percentage of 3 (p=.00194) and 4 (p=.00813) G-repeats and their location in the mRNA (Figure 6D-E). UTR and 5' LG4s were less dense in G-triplets compared to downstream middle and 3' LG4s (Figure 6D). An inverse

relationship was observed for quadruplets of G. UTR-LG4 were especially G-quadruplet rich while the 3' LG4 average percentage of G-quadruplet repeats was reduced (Figure 6E). There was an enrichment (p=.0048) of G-quintuplets only in 5' introns (Figure 6F).

These results mark the first analyses on the quantity of guanine repeats for G4 sequences in relation to their positions within an mRNA. The G4 motif density and %G showed no relationship. This implies that classifying different G4 sequences based on their G-repeat compositions, and not simply how well they form G4 or %G, is necessary to gain a better understanding of the connections between sequence and cellular function.

The precise type of quadruplex that forms is dependent upon the sequences participating in structure formation (Burge et al., 2006), suggesting that within these long G4 sequences there may be multiple types of G4s, and that could translate to different biological functions. This indicates that there could be a relationship between the make up of guanine repeats and potential regulatory function. In fact this seems plausible with promoter LG4s encoding a high G-quintuplet (p=.016) and low G-triplet ratio (p=.011) (Figure 7A, C), coinciding with 5' LG4s containing a high G-quintuplet density (Figure 6F). Open chromatin regions showed an inverse relationship, similar to other LG4s (Figure 7A, C). This suggests that features outside of 3, 4, or 5 G-repeats permit these regions to adopt open chromatin states. There was no significant relationship between the percentage of G-quadruplets and potential for regulation (Figure 7B). The precise sequences potentially composing G4 structures and their

impact on biology is yet to be defined, however the identification of common repeat lengths in promoter regions may help decipher the influence of G4 DNA on gene regulation.

A Subset of LG4s form G-Quadruplex *In Vitro*

While computer programs can predict the G4 folding potential of LG4 sequences, it is important to test the ability for these sequences to form structures experimentally. It is not practical to test every sequence, so a subset was selected for experimental validation. My goal was to verify G4 structure formation in LG4s. Based on differential sequence composition of G-repeats among LG4, 15 LG4s from transcribed regions (10%) were assayed using circular dichroism (CD). CD measures the differential absorption of left and right polarized light from chiral molecules in solution in order to identify structural conformations (Gray and Bollum, 1974). Table 8 shows the name of the LG4 region tested next to each corresponding oligonucleotide sequence. G4 can adopt two different conformations: parallel and anti-parallel, which describes the directionality of the DNA strands composing the structure (Burge et al., 2006). Parallel G4 DNA results in a CD spectrum (called ellipticity) with a peak at ~260 nm and dip at ~240 nm. Anti-parallel G4 structures show a peak at ~295 nm and dip at ~260 nm (Balagurumoorthy et al., 1992; Giraldo et al., 1994). All LG4 oligonucleotides tested had spectra characteristic of parallel G4 (Figure 8). In *F7* LG4, there is evidence of some anti-parallel G4 formation (Red line peak 295, Figure 8). These results are not surprising considering that the high concentration of

oligonucleotide used in CD studies would likely favor the formation of multi-stranded (inter-molecular) G4 (Sen and Gilbert, 1992).

To confirm that the G4 structures identified by CD actually form G4 DNA, a subset of the G4 sequences shown (Table 8) were assayed using a polymerase extension assay. Sequences were TOPO cloned, sequenced, and closed-circular single-stranded templates were generated with M13 helper phage. In these assays, polymerase pausing is K^+ dependent and occurs only when the guanine-rich strand serves as the template (Sun and Hurley, 2010; Weitzmann et al., 1996). For the Sγ3 sequence, extension by Klenow polymerase was blocked in an orientation and K^+ dependent manner, which indicates that G4 formation on the template strand inhibited DNA synthesis (Figure 9A) (Ehrat et al., 2012). Using this same assay, 8 LG4 sequences stalled Klenow extension in K^+ dependent manner, very similar to Sγ3 (Figure 9A). Two LG4s (*P2RX5*-G, *HCN2*-G) stalled Klenow in Li^+, although to a lesser extent (Figure 9B). Since Li^+ ions do not support intra-molecular quadruplex formation (Sen and Gilbert, 1990; Kankia and Marky, 2001), these results indicate that non-G4 structures can form in certain G4 sequences. In order to rule out hairpin formation, the reverse complement was assayed. *P2RX5*-C and *HCN2*-C did not stall Klenow extension in either salt (Figure 9B), indicating that the structure that inhibited synthesis required the guanine strand.

LG4s in Transcribed Protein Regions Show Increased Copy Number Variations

In the human genome, copy number variations (CNVs) are major contributors to genetic diversity and increase susceptibility to a range of different genetic disorders (Feuk et al., 2006; Stankiewicz and Lupski, 2010). CNVs are detected through genome wide sequencing studies, submitted to the dbVAR database, and that data is readily available for analysis on NCBI.org (Lappalainen et al., 2013). Since G4 sequences have been found at some regions of increased genetic instability, LG4s potential effect on large sequence variations was investigated. I calculated the density of LG4 CNV breakpoints and compared that to surrounding transcripts.

CNVs for each LG4 in a transcribed region were downloaded from dbVAR on NCBI.org. The number of CNV breakpoints /1kb was calculated for each LG4, 5 kb 5' and 3' of LG4 in 1 kb increments, as well as the remaining transcript not associated with LG4. CNVs less than 100 bp were removed from the analysis. LG4 regions contained a highly significant (p= 0.00697) 8-fold increase in CNV breakpoints compared to loci >2 kb away and nearby (unrelated) transcripts (Red line LG4 central x-axis vs. dashed green line Figure 10A). Unexpectedly, regions 1 kb 5' and 3' had a ~3 fold increase in CNVs, suggesting that LG4s can invoke instability at proximal sequences (5'-1 and 3'-1 Figure 10A). This supports recently reported data suggesting that DNA structures can induce mutagenesis in surrounding regions, a process known as Repeat Induced Mutagenesis (Shishkin et al., 2009).

Due to differences in G-repeat composition at distinctive locations in the mRNA transcript, the LG4 mRNA location was assayed for its effect on the quantity of CNVs. There were no significant differences between the number of CNVs and location in the LG4 (LG4 central x-axis, Figure 10B). Interestingly, loci positioned 1 kb 3' proximal of UTR-LG4 were significantly (p=0.025) increased for CNV breakpoints (blue line 3' 1 kb Figure 10B). This increase of CNVs 3' proximal of UTR-LG4s coincides with a similar increase in G4 density 3' proximal UTR-LG4s (location of 5' introns) (Figure 4B). This could indicate that smaller, less dense regions of G4 can lead to an increase in copy number variation density, or that UTR-LG4 regions have a propensity to inflict instability in adjacent 3' regions for unknown reasons.

The formation of stable G4 structures on the non-transcribed strand (GGG mRNA) can lead to an increase in gross chromosomal rearrangements in yeast (Kim and Jinks-Robertson, 2012; Yadav et al., 2014). Comparisons between the orientations of the G-triplets with respect to the transcribed strand showed no difference (p=.94) in CNV density between the two (Figure 10C). This indicates repeat composition in the mRNA is not a major contributing factor, if one at all, to the increase in LG4 copy number variations observed.

To determine if other factors outside of mRNA strand composition correlated to an increase of CNVs, LG4's regulatory ability was compared to the density of CNVs. I found that promoter-associated LG4s had a significantly (p=0.02) lower number of CNVs when compared to non-promoter CNVs (Figure 10D). Open chromatin regions were extremely prone to CNVs with a ~2-fold

increase compared to other LG4 regions, and ~3-fold increase compared to promoter LG4s (Figure 10D). Further statistical analysis of CNVs compared to LG4 characteristics such as length, G4 motif density, and repeat composition had no significant relationship. This correlation with regulatory elements could be due to multiple mechanisms; including but not limited to the sequence composition of the regulatory motif increasing instability, evolutionary selection against large deletions or insertions of important regulatory regions, or the different interactions of trans regulatory elements. Whatever the reason, LG4s seem to be unstable and showed an increase in CNV breakpoints compared to the surrounding DNA. Furthermore, the CNV breakpoints extend to sequences that flank the guanine repeats.

Small Insertions and Deletions in Expressed LG4

The mechanisms promoting large chromosome rearrangements are not known. It is believed that aberrant resolution of DNA breaks is responsible in initiating mutagenesis (Kasparek et al., 2011). However, most breaks are repaired in a accurate manner by the homologous recombination (HR) or non homologous end joining (NHEJ) repair pathways. NHEJ repair of DSBs is typically imprecise and frequently leads to small insertions and deletions (Lieber et al., 2003). HR is the predominant repair pathway during replication and was once thought to be error free (Thompson and Schild, 2001). However, translesion polymerase activity during HR repair has demonstrated that this pathway can also induce deletions (Kane et al., 2012). Introduction of mutations during DNA repair is best

exemplified by a study that induced programed DNA breaks with I-SceI endonucleases. They found that small insertions are frequently inserted at DNA break sites, and most likely occur as an alternative to gross chromosomal rearrangements (Onozawa et al., 2014).

If double stranded breaks occur at the largest G4 sequences in the human genome (LG4) at a rate that is higher than surrounding regions, we would predict an increase in small insertions and deletions at those sites. To further study the impact of G4 on DNA breaks and genome instability, the quantity of small insertion and deletion sequence variation events (<100bp) in LG4 sequences compared to surrounding introns was analyzed. Using data from genome wide sequences studies on Ensembl69 (dbSNP release 138 database) (Sherry et al., 2001), the number of insertions and deletions for each LG4 found in transcribed regions was calculated.

The average number of small insertions and deletions per 100 base pairs was significantly increased in LG4 (p=1.12X 10^{-37} and p=1.65 X 10^{-68} respectively) compared to surrounding intronic regions (Figure 11). Deletions had the largest increase (~6-fold) in LG4 (Figure 11A). Insertions in LG4 were not as pronounced, but were present 2.5-fold more than the surrounding intronic regions (Figure 11B). Although an increase of sequence variations in both databases cannot confirm that double stranded breaks occur at LG4s, such a high increase of insertions, deletions, and CNVs supports a model whereby LG4 sequences promote site-specific mutagenesis. Further experimental analysis is needed to

determine if small insertions and deletions, as well as CNVs, are from the repair of double stranded breaks that were triggered by LG4 motifs.

Evidence for LG4 Repeat Length Polymorphisms

Recently, next generation sequencing advances resulted in identification of a G4-repeat expansion in C9ORF72, which induces ALS (Renton et al., 2011; Haeusler et al., 2014). To date, this is the only documented example of a G4 repeat expanding in length. However, one would predict that other G4 motifs in the human genome are also subject to expansion and contraction. In trinucleotide repeat expansion, the longer the repeat, the more prone it is to expansion (McMurray, 2010). This suggests that larger G4 sequences may have emerged over evolutionary time as a product of a repeat expansion mechanisms, or will be polymorphic in general. In support of this hypothesis, many CNVs that I found at LG4s were the result of large duplications of the repetitive unit, suggesting that genome wide sequencing studies have already detected multiple repeat expansions in LG4 regions.

Subsequent experimental evidence for LG4 expansion has also been detected for one LG4 region. During PCR amplification from human genomic DNA, Cytokine Receptor-Like Factor 2 (*CRLF2*) contained multiple bands (Figure 12). The primary PCR product (600 bp) was consistent with the reported LG4 size in the latest human genome release (ch38, 2014). I hypothesized that this was potentially a larger *CRLF2* G4 repeat present in the human population from G4 expansion. A subsequent blast of the primers used displayed very high

specificity the sequences flanking *CRLF2* LG4 (other blast hit E-values >0.1). Further analysis of potential off target primer binding sites did not correspond to the observed increase in PCR product size.

To determine if the larger *CRLF2* LG4 repeat exists in the human genome, the 1,100 bp PCR product (red arrow, Figure 12) was gel excised, TOPO cloned and Sanger sequenced with both forward and reverse primers. The sequencing results indicated that a larger *CRLF2* LG4 intron could be present in the human genome. Subsequent genome searches verified that a G4 motif with this sequence is not present outside of *CRLF2*, indicating that the large PCR product was in fact from the *CRLF2* LG4 sequence in question. Interestingly, the larger 1,100 bp LG4 size was reported in the previous genome release (hg18, 2009), hence its original detection with our Java G4 search program. It is also possible that the repeat expansion can be simply an artifact from PCR. This would also be exciting because it would indicate that G4 could expand through simple replication cycling. Inquiry into G4 expansion in *CRLF2* is warranted due to its similarity to *C9ORF72*, a gene involved in hereditary ALS. These similarities include G-rich intronic mRNA, 3' transcribed region, potential neurological disease involvement, and a repetitive unit capable of G4 formation (Haeusler et al., 2014, own analysis). First however, verification of the sequence of the smaller PCR product is needed to see if it can expand to the larger product in PCR reactions, and if this region forms stable G4 structures.

Intronic LG4s and Human Disease

LG4s are found in multiple human diseases, meriting additional analysis. Through extensive literature searches and use of the DAVID web interface (Huang et al., 2008), LG4s were found in over 27 cancer related proteins, 16 proteins involved in developmental diseases, and 18 proteins involved in neurological diseases (Table 9-11). In total, 64/154 (41%) of the LG4s are involved somehow in human disease. In comparison, genecards.org lists 6,548/ 38,656 (16%) of human proteins are involved in human disease (Safran et al., 2010).

Discussion

To review the major findings of this chapter, a new set of large G4 regions found outside of immunoglobulin switch regions exist in the human genome. These results support a hypothesis in which sequences containing G4 repeats promote mutagenesis. This is the first study to specifically connect large G4 motifs to sequence variation, as well as identify multiple prominent candidates for site-specific genetic instability from G4 structure formation. Not only have LG4s been shown to house a large increase in sequence variations, implying they are genetically unstable, but their presence in regulatory genes highlight an importance to understanding the connections between G4 structures and site-specific instability.

The plasticity of the LG4 identification program allows for the sliding window size to be reduced while still maintaining a similar G-triplet/sequence length ratio threshold used in the current search. This would be useful for

highlighting multiple G4 transcribed regions that are smaller in size (200-600 bp), which may still be biologically relevant. Another avenue of study is to determine if smaller G4 motifs are located in regions showing increased genome instability, or if this is a feature of only the largest G4 sequences in the human genome.

Additional analyses revealed that the LG4 loci cannot be classified simply as able to form G4, how “G-rich” they are, or density of G4 motifs. This is because there was a strong relationship between composition of repeats, location in the mRNA, and potential regulatory function. The difference is most likely due to sequence-specific regulatory properties from alternate structure conformations, or to undefined sequence requirements for trans acting regulatory elements. One inference drawn from these data is that extensively repetitive guanine repeats form highly stable G4 structures. This is especially prevalent in UTR and 5’ regions of the mRNA, which contained longer G-repeats on average. That may explain why many LG4 are found in promoter regions. The regulatory contributions for internal and 3’ G4 motifs remain unclear. Interestingly, internal and 3’ G4 sequences had higher G-triplet repeat compositions compared to LG4 in 5’ UTR and promoter regions. From this one can deduce that alternate G4 structures or stability in those regions translates to some other, yet to be defined, function. Investigating characteristics of middle and 3’ LG4 could provide insights into deciphering their regulatory function.

One important question raised by this investigation is the link between LG4 regulatory function and the density of copy number variations (Figure 10D). It is possible that only certain types of G4 motifs are unstable because of their

genomic location. For instance, promoter LG4s may lead to lower levels of mutagenesis because of the specific type of G4 structure found in promoter regions. However, promoter regions technically form more stable G4 motifs (longer G-repeats), and one would predict that the most stable G4 structures would lead to the most instability. An increase in stability of G4 structures may increase the potential for mutagenesis. This is supported by *in vivo* studies where stabilization of G4 by ligands, or increase in repeat size, amplified mutagenesis at those sequences (Piazza et al., 2010; Ribeyre et al., 2009).

I favor a hypothesis that interprets these results through an evolutionary point of view. Due to the necessity and stringency of certain regulatory mechanisms and their corresponding sequence compositions, it is possible that promoter associated LG4 regions are prone to the same instability found in 3' mRNA regions. In spite of this, any duplication or deletion in LG4 promoter length interrupts its regulatory ability and is severely selected against by evolutionary forces. In support of that idea, non-promoter open chromosomal regions have an extremely large increase in sequence variations. It is feasible that the regulatory mechanisms in middle and 3' regions are highly susceptible to sequence alterations without negative selection, and in turn display a high rate of variation in databases. It is also possible that an open chromosomal state from G4 formation leads to unprotected DNA and increased mutagenesis from both molecular and chemical DNA damaging agents (Roberts et al., 2012). Further research is needed to link LG4 repeat make up with regulatory ability and the

inherent instability of those sequences. Most likely, multiple interlinking factors are involved, making this a complicated issue.

Although our search focused on identifying regions of G4 formation, multiple LG4s are capable of other non-B form DNA structures. Further computational analysis revealed multiple regions capable of stable hairpin formation and long purine repeats (ex AGGGA) capable of triplex structures (Frank-Kamenetskii and Mirkin, 1995). Perhaps the most interesting find was that a subset of cloned LG4 motifs stalled Klenow polymerase in a K^+ independent manner, suggesting other structures formed that stalled DNA synthesis. Although it was observed that stalling was still increased in K^+ (Figure 9B) and that CD scans demonstrated parallel G4 formation occurs (Figure 8), the results suggest the presence of (non-G4) structures. Hairpins can also stall replication (Voineagu et al., 2008). Even so, stalling was only observed when using a G-rich template, suggesting the presence of hairpins in the complement would have a similar capacity to form those structures. Also, LG4 sequences that showed stalling in Li^+ did not contain any long stretches of purines, so it seems unlikely that this could be due to triplex DNA. K^+ independent stalling was also reported in telomeres, and to date, any details of an alternative structure to G4 formation remain illusive (Lormand et al., 2013). It is also feasible that subsets of intra-molecular G4 structures do not require K^+ to form.

G4 repeat expansion is a relatively new phenomenon that has major implications in human disease. It is believed that *C9ORF72* repeat expansion leads to ALS from an increase in aborted transcripts at an expanded G4-mRNA

intron, leading to nuclear coagulation of aborted transcripts and ribonucleoproteins that induces nerve cell stress, neurodegenerative damage, and a diseased phenotype (Haeusler et al., 2014). Therefore, expansion of any mRNA containing a large G4 sequence could have similar implications, especially in neurological disorders. Outside of *CRLF2* variations in LG4 length, many LG4 loci with high CNV breakpoint density contained large duplications of the G4 repeat. For example, CNV analysis of Transmembrane Protease Serine 2 (*TMPRSS2*) on Ensembl77 reveals that large duplications of its LG4 intron have been detected 25 times in genome wide sequencing studies. Considering *TMPRSS2* is overexpressed in a majority of prostate carcinomas (Vaarala et al., 2001), the effect of LG4 on *TMPRSS2* regulation and site-specific instability should be examined.

Previous *in vivo* studies of G4-instigated gross chromosomal rearrangements have demonstrated mutagenesis is drastically increased when the G-rich sequence is on the non-transcribed strand compared to the opposite orientation (Yadav et al., 2014). To my surprise there was no difference between CNV density and LG4 orientation with respect to transcriptional orientation. This could be due to different sequence compositions of the repetitive units tested. The previous study by Yadav et al., 2014 used a G4 sequence found in murine switch regions that is similar to our model G4 sequence, Sγ3. Switch regions are less dense in G-triplets and were used as the minimum definition of G-triplet density in this study. Therefore, LG4s on average contain shorter loops and a higher density of G-repeats, classic definitions of more “stable” G4 (Burge et al.,

2006). With this in mind, it is possible that transcriptional orientation has no impact in mutagenesis after a certain size or density threshold is reached.

This study has compiled a comprehensive database of large G4 regions in the human genome consisting of their specific characteristics such as sequence composition, length, location, small sequence variations, large sequence variations, regulatory capability, correlation with human disease, and ability to form other structures. This has allowed me to connect genome instability with G4 sequences. Future *in vivo* analysis should investigate the effects individual LG4 loci have on gross chromosomal rearrangements. It would be beneficial to start with LG4s containing a high density of CNVs and located in proteins that lead to disease. This could identify areas of site-specific instability instigated by G4 that subsequently lead to disease. Some examples are *CTCFL*, *ANO9*, *PRAME*, *SARDH*, *P2RX5*, *TCF3*, and *ABR*. It would also be interesting to assay LG4s containing a low number of CNVs to see if these regions can also lead to instability while remaining undetected in genome wide sequencing studies.

Upon identification of extensively repetitive G4 sequences in the human genome (LG4), the additional analysis presented in this chapter further characterized these regions to clarify the spectrum of sequence changes at those loci and roles in regard to gene regulation. This same methodology can be applied to other organisms, or used to compare G4 motifs across genomes. Sequence alignments revealed that 7/7 of the LG4s are primate specific motifs, but LG4 length was shorter in other primates (not shown). While G4 is not confined to the human genome, it would be interesting to determine whether the

LG4 sequences are conserved among species. If LG4 size was conserved across Eukaryotes, it would imply the length of G4 motifs is important in regulation. However, preliminary results suggest that the large length is restricted to primates. Smaller, or complete lack of G4 motifs outside of primates would suggest that our use of G4 has evolved in complexity. Considering multiple G4 motifs are found in developmental and cell cycle regulatory proteins, these differences could be part of what defines us as a species and provide insights into our evolutionary past.

References

Arora, A., Suess, B., (2011). An RNA G-quadruplex in the 3'UTR of the proto-oncogene PIM1 represses translation. *RNA biology*, 8:802-805.

Balagurumoorthy, P., Brahmachari, S. K., Mohanty, D., Bansal, M., & Sasisekharan, V. (1992). Hairpin and parallel quartet structures for telomeric sequences. *Nucleic Acids Research*, 20(15), 4061-4067.

Baral, A., Kumar, P., Halder, R., Mani, P., Yadav, V.K., Singh, A., Chowdhury, S. (2012). Quadruplex-single nucleotide polymorphisms (Quad-SNP) influence gene expression difference among individuals. *Nucleic Acids Research* 40: 3800-3811.

Beaudoin, J. D., & Perreault, J. P. (2013). Exploring mRNA 3′-UTR G-quadruplexes: evidence of roles in both alternative polyadenylation and mRNA shortening. *Nucleic Acids Research*, 41(11), 5898-5911.

Booth, H. A. F., & Holland, P. W. (2007). Annotation, nomenclature and evolution of four novel homeobox genes expressed in the human germ line. *Gene* 387:7-14.

Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K., & Neidle, S. (2006). Quadruplex DNA: sequence, topology and structure. *Nucleic acids research*, 34(19), 5402-5415.

Capra, J. A., Paeschke, K., Singh, M., & Zakian, V. A. (2010). G-quadruplex DNA sequences are evolutionarily conserved and associated with distinct genomic features in Saccharomyces cerevisiae. *PLoS computational biology*, 6(7), e1000861.

Cogoi, S., Xodo, L.E. (2006). G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic Acids Res*earch 34:2536-2549.

Du, X., Gertz, E. M., Wojtowicz, D., Zhabinskaya, D., Levens, D., Benham, C. J., & Przytycka, T. M. (2014). Potential non-B DNA regions in the human genome are associated with higher rates of nucleotide mutation and expression variation. *Nucleic acids research*, 42(20), 12367-12379.

Duquette, M. L., Handa, P., Vincent, J. A., Taylor, A. F., & Maizels, N. (2004). Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. *Genes & development*, 18(13), 1618-1629.

Eagle, L.R., Yin, X., Brothman, A.R., Williams, B.J., Atkin, N.B., Prochownik, E.V. (1995). Mutation of the MXI1 gene in prostate cancer. *Nature genetics* 9:249-255.

Eddy, J., Maizels, N. (2008). Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. *Nucleic Acids Res*earch 36:1321-1333.

Ehrat, E.A., Johnson, B.R., Williams, J.D., Borchert, G.M., Larson, E.D. (2012). G-quadruplex recognition activities of E. Coli MutS. *BMC molecular biology* 13: 23.

Elgar, G., Vavouri, T. (2008). Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends in genetics* 24:344-352.

Feuk, L., Carson, A.R., Scherer, S.W. (2006). Structural variation in the human genome. *Nature Reviews Genetics* 7:85-97.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Hubbard, T. J. (2011). Ensembl 2012. *Nucleic Acids Res*earch, gkr991.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Yates, A. (2013). Ensembl 2014. *Nucleic Acids Res*earch gkt1196.

Folle, G. A. (2008). Nuclear architecture, chromosome domains and genetic damage. *Mutation Research/Reviews in Mutation Research*, 658(3), 172-183.

Frank-Kamenetskii, M.D., Mirkin, S.M. (1995). Triplex DNA structures. *Annual review of biochemistry* 64:5-95.

Gellert, M., Lipsett, M.N., Davies, D.R. (1962). Helix formation by guanylic acid. *Proceedings of the National Academy of Sciences of the United States of America* 48:2013.

Giraldo, R., Suzuki, M., Chapman, L., Rhodes, D. (1994). Promotion of parallel DNA quadruplexes by a yeast telomere binding protein: a circular dichroism study. *Proceedings of the National Academy of Sciences*, 91:7658-7662.

Gray, D.M., Bollum, F.J. (1974). A circular dichroism study of poly dG, poly dC, and poly dG: dC. *Biopolymers* 13:2087-2102.

Haeusler, A.R., Donnelly, C.J., Periz, G., Simko, E.A., Shaw, P.G., Kim, M.S., Wang, J. (2014). C9orf72 nucleotide repeat structures initiate molecular cascades of disease. *Nature* 507:195-200.

Huang, D.W., Sherman, B.T., Lempicki, R.A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* 4:44-57.

Huppert, J.L., Balasubramanian, S. (2005). Prevalence of quadruplexes in the human genome. *Nucleic Acids Res*earch 33:2908–2916.

Huppert, J.L., Balasubramanian, S. (2007). G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res*earch 35:406-413.

Kane, D.P., Shusterman, M., Rong, Y., McVey, M. (2012). Competition between replicative and translesion polymerases during homologous recombination repair in Drosophila. *PLoS genetics* 8:e1002659.

Katapadi, V.K., Nambiar, M., Raghavan, S.C. (2012). Potential G-quadruplex formation at breakpoint regions of chromosomal translocations in cancer may explain their fragility. *Genomics* 100:72-80.

Kankia, B. I., & Marky, L. A. (2001) Folding of the thrombin aptamer into a G-quadruplex with Sr2+: stability, heat, and hydration. Journal of the American Chemical Society, 123(44), 10799-10804.

Kasparek, T. R., & Humphrey, T. C. (2011) DNA double-strand break repair pathways, chromosomal rearrangements and cancer. In Seminars in cell & developmental biology (Vol. 22, No. 8, pp. 886-897). Academic Press.

Kersey, P. J., Allen, J. E., Christensen, M., Davis, P., Falin, L. J., Grabmueller, C., and Staines, D. M. (2014). Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res*earch, 42(D1), D546-D552.

Kikin, O., D'Antonio, L., Bagga, P.S. (2006). QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res*earch 34:W676-W682.

Kim, N., Jinks-Robertson, S. (2012). Transcription as a source of genome instability. *Nature Reviews Genetics* 13:204-214.

Koole, W., van Schendel, R., Karambelas, A.E., van Heteren, J.T., Okihara, K.L., Tijsterman, M. (2014). A Polymerase Theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites. *Nature communications* 5.

Kumar, N., Basundra, R., Maiti, S. (2009). Elevated polyamines induce c-MYC overexpression by perturbing quadruplex–WC duplex equilibrium. *Nucleic Acids Res*earch 37:3321-3331.

Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., ... & Church, D. M. (2013). DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res*earch 41:D936-D941.

Laud, P. R., Multani, A. S., Bailey, S. M., Wu, L., Ma, J., Kingsley, C., ... & Chang, S. (2005). Elevated telomere-telomere recombination in WRN-deficient, telomere dysfunctional cells promotes escape from senescence and engagement of the ALT pathway. *Genes & development* 19:2560-2570.

Lieber, M. R., Ma, Y., Pannicke, U., & Schwarz, K. (2003). Mechanism and regulation of human non-homologous DNA end-joining. *Nature reviews Molecular cell biology*, *4*(9), 712-720.

Lormand, J. D., Buncher, N., Murphy, C. T., Kaur, P., Lee, M. Y., Burgers, P., Opresko, P. L. (2013). DNA polymerase δ stalls on telomeric lagging strand templates independently from G-quadruplex formation. *Nucleic Acids Res*earch 41:10323-10333.

Maizels, N. (2006). Dynamic roles for G4 DNA in the biology of eukaryotic cells. Nature structural & molecular biology 13:1055-1059.

Maizels, N. (2012). G4 motifs in human genes. A*nnals of the New York Academy of Sciences* 1267: 53-60.

Marcel, V., Tran, P. L., Sagne, C., Martel-Planche, G., Vaslin, L., Teulade-Fichou, M. P., Van Dyck, E. (2011). G-quadruplex structures in TP53 intron 3: role in alternative splicing and in production of p53 mRNA isoforms. Carcinogenesis, 32(3), 271-278.

McMurray, C.T. (2010). Mechanisms of trinucleotide repeat instability during human development. *Nature Reviews Genetics* 11:786-799.

Nakken, S., Rognes, T., Hovig, E. (2009). The disruptive positions in human G-quadruplex motifs are less polymorphic and more conserved than their neutral counterparts. *Nucleic Acids Research* gkp590.

Nambiar, M., Goldsmith, G., Moorthy, B. T., Lieber, M. R., Joshi, M. V., Choudhary, B., Raghavan, S. C. (2011). Formation of a G-quadruplex at

the BCL2 major breakpoint region of the t (14; 18) translocation in follicular lymphoma. *Nucleic Acids Res*earch 39:936-948.

Nambiar, M., Srivastava, M., Gopalakrishnan, V., Sankaran, S. K., & Raghavan, S. C. (2013). G-quadruplex structures formed at the HOX11 breakpoint region contribute to its fragility during t (10; 14) translocation in T-cell leukemia. *Molecular and cellular biology*, 33(21), 4266-4281.

Onozawa, M., Zhang, Z., Kim, Y. J., Goldberg, L., Varga, T., Bergsagel, P. L., Aplan, P. D. (2014). Repair of DNA double-strand breaks by templated nucleotide sequence insertions derived from distant regions of the genome. *PNAS*, *111*(21), 7729-7734.

Rawal, P., Kummarasetti, V. B. R., Ravindran, J., Kumar, N., Halder, K., Sharma, R., and Chowdhury, S. (2006). Genome-wide prediction of G4 DNA as regulatory motifs: role in Escherichia coli global regulation. *Genome research*, *16*(5), 644-655.

Ribeyre, C., Lopes, J., Boulé, J. B., Piazza, A., Guédin, A., Zakian, V. A., Nicolas, A. (2009). The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. *PLoS genetics,* 5(5), e1000475.

Roberts, S. A., Sterling, J., Thompson, C., Harris, S., Mav, D., Shah, R., ... & Gordenin, D. A. (2012). Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions *Molecular cell* 46:424-435.

Renton, A. E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J. R., Sulkava, R. (2011). A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron*, *72*(2), 257-268.

Piazza, A., Boulé, J. B., Lopes, J., Mingo, K., Largy, E., Teulade-Fichou, M. P., and Nicolas, A. (2010). Genetic instability triggered by G-quadruplex interacting Phen-DC compounds in Saccharomyces cerevisiae. *Nucleic Acids Res*earch, 38(13), 4337-4348.

Safran, M., Dalah, I., Alexander, J., Rosen, N., Stein, T. I., Shmoish, M., ... & Lancet, D. (2010). GeneCards Version 3: the human gene integrator. *Database 2010*baq020.

Samuel, S., Naora, H. (2005). Homeobox gene expression in cancer: insights from developmental regulation and deregulation. *European Journal of Cancer* 41:2428-2437.

Sen, D., Gilbert, W. (1988). Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature* 364-366.

Sen, D., Gilbert, W. (1990). A sodium-potassium switch in the formation of four-stranded G4-DNA. *Nature* 344:410-414.

Sen, D., Gilbert, W. (1992). Novel DNA superstructures formed by telomere-like oligomers. Biochemistry 31:65-70.

Sen, D., & Gilbert, W. (1992). Guanine quartet structures. *Methods in enzymology*, *211*, 191-199.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*earch, 29(1), 308-311.

Shishkin, A. A., Voineagu, I., Matera, R., Cherng, N., Chernet, B. T., Krasilnikova, M. M., . Mirkin, S. M. (2009). Large-scale expansions of Friedreich's ataxia GAA repeats in yeast. *Molecular cell* 35:82-92.

Siddiqui-Jain, A., Grand, C. L., Bearss, D. J., & Hurley, L. H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proceedings of the National Academy of Sciences* 99:11593-11598.

Stankiewicz, P., Lupski, J.R. (2010). Structural variation in the human genome and its role in disease. *Annual review of medicine* 61:437-455.

Subramanian, M., Rage, F., Tabet, R., Flatter, E., Mandel, J. L., & Moine, H. (2011). G–quadruplex RNA structure as a signal for neurite mRNA targeting. *EMBO reports* 12:697-704.

Sun, D., Guo, K., Rusche, J.J., Hurley, L.H. (2005). Facilitation of a structural transition in the polypurine/polypyrimidine tract within the proximal promoter region of the human VEGF gene by the presence of potassium and G-quadruplex-interactive agents. *Nucleic Acids Res*earch 33:6070-6080.

Thompson, L. H., & Schild, D. (2001). Homologous recombinational repair of DNA ensures mammalian chromosome stability. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 477(1), 131-153.

Sun, D., & Hurley, L. H. (2010). Biochemical techniques for the characterization of G-quadruplex structures: EMSA, DMS footprinting, and DNA polymerase stop assay. In G-Quadruplex DNA H*umana Press.*

Tarsounas, M., Tijsterman, M. (2013).Genomes and G-quadruplexes: for better or for worse. *Journal of molecular biology* 425:4782-4789.

Todd, A.K., Johnston, M., Neidle, S. (2005). Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res* 33:2901–2907.

Vaarala, M.H., Porvari, K., Kyllönen, A., Lukkarinen, O., Vihko, P. (2001). The TMPRSS2 gene encoding transmembrane serine protease is overexpressed in a majority of prostate cancer patients: detection of mutated TMPRSS2 form in a case of aggressive disease. *International journal of cancer* 94:705-710.

van Kregten, M., Tijsterman, M. (2014). The repair of G-quadruplex-induced DNA damage*. Experimental cell research* 329:178-183.

Voineagu, I., Narayanan, V., Lobachev, K.S., Mirkin, S.M. (2008). Replication stalling at unstable inverted repeats: interplay between DNA hairpins and fork stabilizing proteins. *PNAS* 105:9936-9941.

Weitzmann, M. N., Woodford, K. J., & Usdin, K. (1996). The development and use of a DNA polymerase arrest assay for the evaluation of parameters affecting intrastrand tetraplex formation. *Journal of Biological Chemistry*, 271(34), 20958-20964.

Wu, Y, Brosh, R.M. (2010). G-quadruplex nucleic acids and human disease. *FEBS journal* 277:3470-3488.

Yadav, P., Harcy, V., Argueso, J.L., Dominska, M., Jinks-Robertson, S., Kim, N. (2014). Topoisomerase I Plays a Critical Role in Suppressing Genome Instability at a Highly Transcribed G-Quadruplex-Forming Sequence. *PLoS genetics* 10:e1004839.

Table 1. Location of Large G4 Capable Regions (LG4s) in the Human Genome Relative to Transcription.

Over 315 individual LG4s exist in the human genome. They were found in transcribed regions (Transcribed), regions unassociated with transcription (Unassociated >25 kb), 0-2 kb adjacent to the UTR (<2 kb 5' UTR) and 2-25 kb from the closest UTR (2-25 kb UTR). Number of individual LG4 loci for each location is listed (number loci). The percentage of each LG4s location compared to total number (%).

LG4 location	number loci	%
Transcribed	154	49%
<2 kb 5' UTR	26	8%
2-25 kb UTR	34	11%
unassociated >25kb	101	32%
total	315	

Table 2. mRNA Location of LG4s.
LG4s were found in UTRs, Introns, and one exon (LG4 location). The UTR/intronic locations of LG4s varied and are found in all parts of the transcript. In the first column, three locations are listed (UTR, intron, and exon in black bold). 5' and 3' sequences within each location are indicated (grey font). The numbers of LG4 loci located at each corresponding mRNA location are listed (number loci). The right column shows the number of LG4s found in each mRNA, the location was dived by the total transcribed LG4 (%).

LG4 location	number loci	%
UTR	**10**	**6%**
5'	9	
3'	1	
intron	**143**	**93%**
5'	62	
middle	49	
3'	32	
exon	**1**	**0.6%**
total	**154**	

Table 3. Guanine or Cytosine Repeats Compose the mRNA Transcript. Sequences supportive of G4 structures were found on both the non-transcribed strand (GGG-mRNA) and transcribed strand (CCC=mRNA). The percent of each mRNA strands composition was calculated (%).

LG4 location	number loci	%
GGG mRNA	62	40%
CCC mRNA	92	60%
total	154	

Table 4. Transcribed LG4s are Located at Regulatory Motifs.
LG4s were found at known regulatory motifs on Ensembl77 Regulatory Build (LG4 regulation). A subset of regulatory motifs contained ChipSeq pull down interactions (LG4 Chip-seq). Some LG4 sequences had multiple regulatory elements present (total regulatory elements). The percentage of LG4s involved in regulation compared to total transcribed LG4s (%).

LG4 on Regulatory Build	number loci	%
LG4 regulation	79	52%
LG4 Chip-Seq	36	
total Regulatory elements	103	
LG4 no regulation	74	48%
total	153	

Table 5. Type of Regulatory Element and its Corresponding mRNA Position.

Distribution of regulatory elements found at LG4 sequences using Ensembl77 Regulatory build. Each type of regulatory motif (regulatory motif) and the total number found (Total) are displayed. The types of regulatory functions identified are: OC-open chromatin, promoter-promoter or promoter flanking regions, TFBS-transcription factor binding site, Enhancer-transcription enhancer motif, CTCF-CCCTC-Binding factor. The type of regulatory element and number found in each mRNA location are displayed and compose the total (UTR, 5', Mid, 3', Exon).

regulatory motif	total	UTR	5'	mid	3'	exon
OC	29	1	7	15	5	1
promoter	36	5	22	5	4	0
TFBS	11	1	3	2	5	0
enhancer	9	0	5	2	2	0
CTCF	18	1	5	6	6	0

Table 6. Transcription Factors Interacting with LG4s.
There were 72 total ChipSeq data entries out of the 36 LG4s that interact with transcription factors indicating multiple transcription factors can interact at a single LG4 locus. 31/36 LG4s, and 31/72 total ChipSeq interactions were with EGR1.

TF pulled down	# of LG4
CTCF	1
E2F4	3
E2F6	3
EBF1	3
EGR1	31
ELF1	1
ETS1	2
FOSL1	2
FOSL2	4
FOXA1	1
JUND	2
MAX	1
Nrf1	1
SP1	7
SRF	1
Tr4	1
USF1	4
Yy1	1
ZBTB33	2
znf263	1
Total	72

Table 7. LG4s Bound by *EGR1* or *SP1* and their Characteristics
(7.1) The type of LG4 regulatory interaction (OC, promoter, TFBS, enhancer, CTCF) and subsequent transcription factor interaction (EGR1, SP1). **(7.2)** LG4 locations in the mRNA with subsequent type of transcription factor interaction (EGR1, SP1).

Table 7.1

TF	OC	promoter	TFBS	enhancer	CTCF
EGR1	3	19	5	2	2
SP1	0	7	0	0	0

Table 7.2

	UTR	5'	mid	3'	exon
EGR1	6	16	2	4	0
SP1	2	5	0	0	0

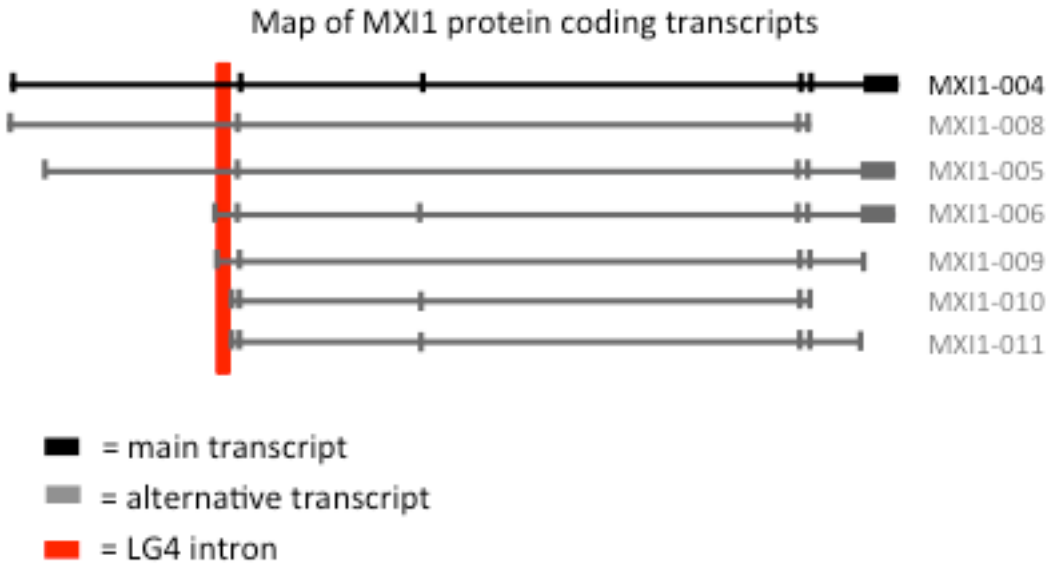

Figure 3. Map of Human MXI1 Main and Alternative Transcripts. The MXI1 oncogene protein coding transcripts are aligned below 5' to 3', with exons/UTRs represented by vertical lines and introns by horizontal lines. The main transcript (MXI1-004) is depicted on top (black lines), while alternative transcripts are depicted below (gray lines). The location of MXI1's LG4 intron is shown (solid red vertical box) and corresponds to multiple alternative transcripts 5' UTR or first intron.

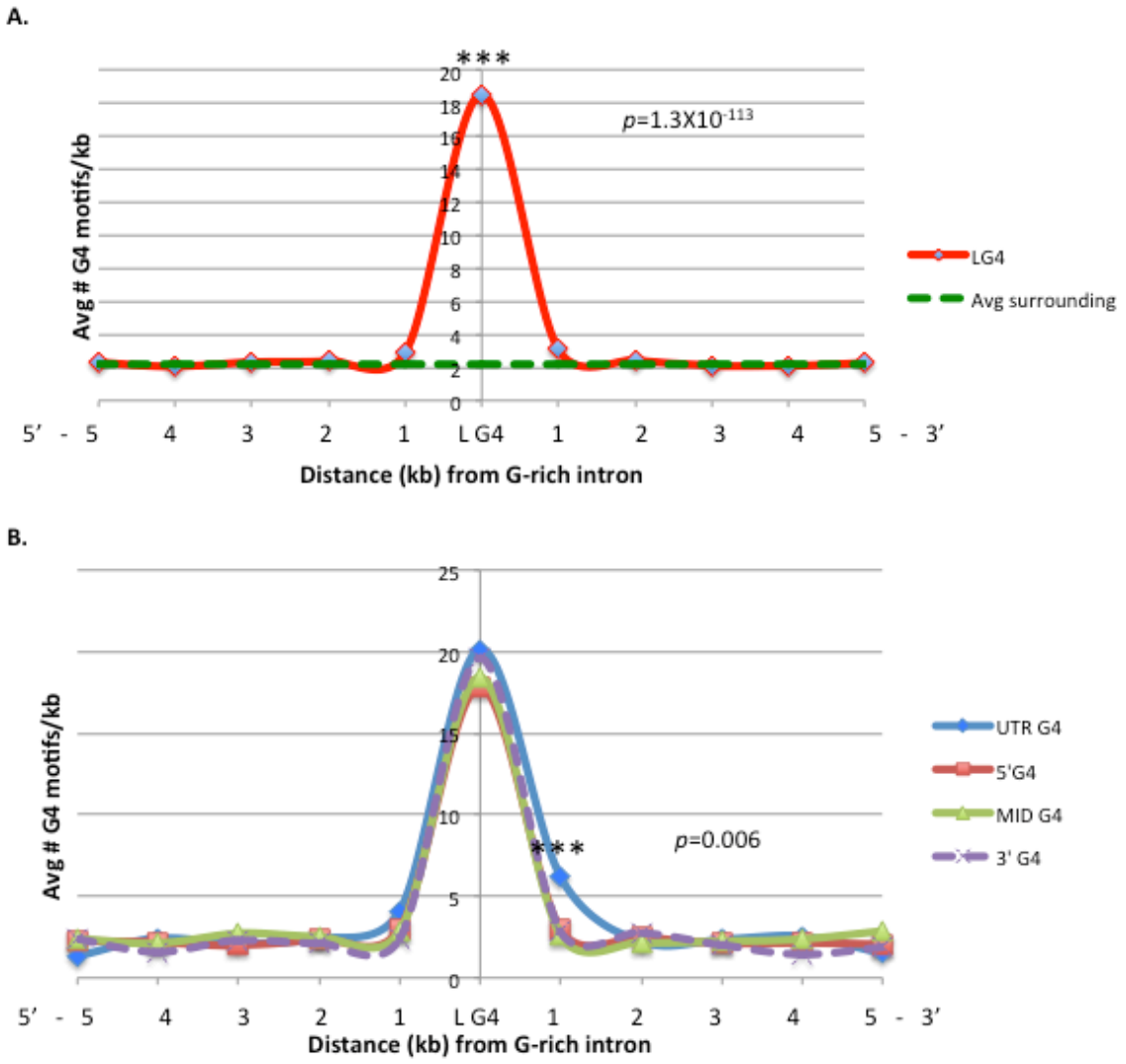

Figure 4. Density of G4 Motifs in LG4s Compared to Surrounding Regions. All graphs display the average number of non-overlapping G4 motifs predicted by QGRS mapper per kb (G4 motif density) (Y-axis) found in LG4s (LG4 central X-axis), and in one kb increments 5' and 3'. **(A)** Density of G4 motifs is increased 9 fold in LG4s (Red line LG4 central X-axis) compared to regions directly flanking both 5' and 3' (5'-1, 3'-1), and the average of the transcript not directly flanking LG4s (average surrounding, green line). **(B)** There average density of G4 motifs and corresponding LG4 mRNA location (LG4 central x-axis) is graphed. UTR-LG4s (blue line) G4 density was statistically increased 1 kb 3' of the LG4 compared to other LG4 mRNA locations.

A.

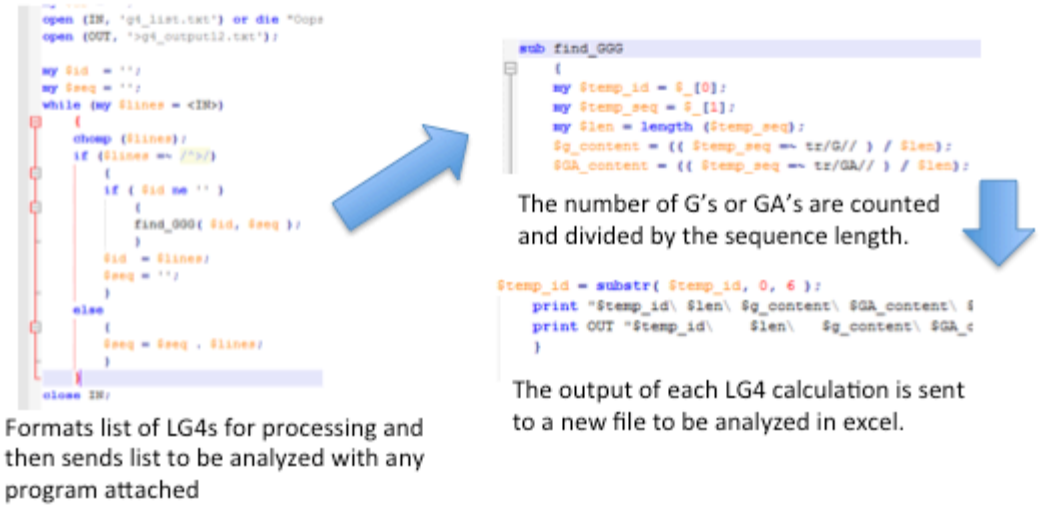

Figure 5A. Perl Programs Used to Count the Sequence Composition of LG4 Repeats. Perl scripts used to count each individual LG4 guanine composition and length. **(A)** First Perl program takes list of LG4s in G-rich FASTA format and formats the list for computational analysis by any desired attached program (sub find_GGG). The first sub program in (A right) counts the total LG4 length and number of G's. The length, and % guanine content for each LG4 is output for analysis in excel.

B.

```
sub find_GGG
    {
    my $temp_id = $_[0];
    my $temp_seq = $_[1];
    my $len = length ($temp_seq);
    $GGG_content = ( $temp_seq =~ s/GGG/X/g);
    $GGGG_content = ( $temp_seq =~ s/XG/Y/g);
    $GGGGG_content = ( $temp_seq =~ s/YG/Z/g);
    $G6_content = ( $temp_seq =~ s/XX/W/g);
    $G2_content = ( $temp_seq =~ s/GG/P/g);
    $Gtrip = ($temp_seq =~ tr/X//);
    $Gquad = ($temp_seq =~ tr/Y//);
    $G5 = ($temp_seq =~ tr/Z//);
    $G6 =  ($temp_seq =~ tr/W//);
    $G2 =  ($temp_seq =~ tr/P//);

    $temp_id = substr( $temp_id, 0, 6 );
        print "$temp_id\ $len\ $g_content\ $GA_content\ $Gtrip\ $Gquad\ $G5\    $G6\    $G2\n";
        print OUT "$temp_id\    $len\    $g_content\ $GA_content\    $Gtrip\ $Gquad\ $G5\    $G6\    $G2\n";
        )
```

Changes the different types of G repeats to countable characters

Counts the number of characters representing each type of G repeat

The output of each LG4 calculation is sent to a new file to be analyzed and calculated in excel.

C.

Visualization of what program B does to count the type of G-run

Figure 5B. Perl Programs Used to Count the Sequence Composition of LG4 Repeats Perl scripts used to count each individual LG4 guanine composition and length. **(B)** Second sub-program that can be attached to first data formatting step in (A left). This sub-program contains multiple steps that changes each length of G-repeat into a countable character (Ex GGG→X, GGGG→Y), counts the characters, and outputs the number counted of each character and LG4 length to an Excel program in order to calculate the individual guanine repeat composition percentage of each longer repeat. **(C)** Visualization of what sub program B does. LG4 in FASTA format (left) consist of multiple lengths of G-repeats, with each color identifying a unique length of G-repeat (middle). This program then identifies G-repeat length and assigns a character that is then counted and number output in excel.

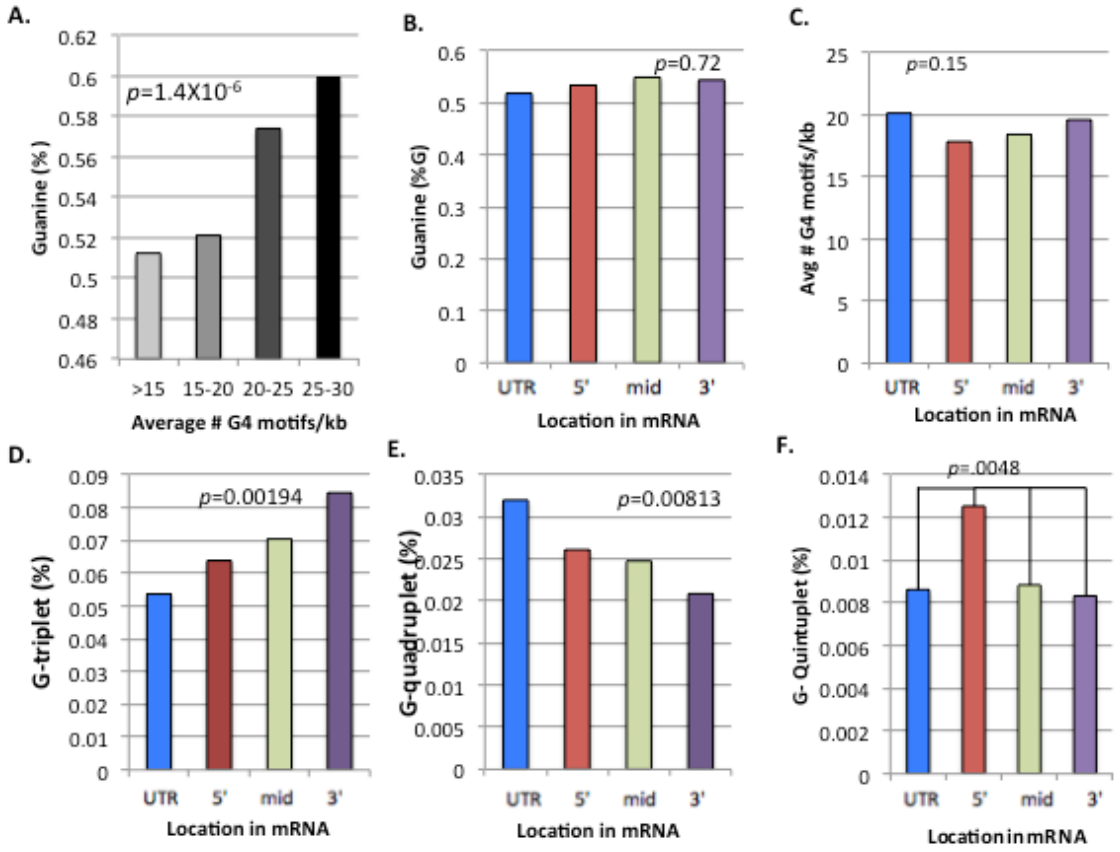

Figure 6. Correlation of Sequence with Location in mRNA. Outputs of Perl programs above were used to identify relationships between LG4 sequences and repeat composition. **(A)** The percentage of guanines (y-axis) and the density of G4 motifs/kb (x-axis) is graphed. **(B)** The percentage of guanine (y-axis) and the location of the LG4 in the mRNA (x-axis) is shown. **(C)** In a similar fashion, the density of G4 motifs/kb (y-axis) is graphed with the LG4 location in the mRNA (x-axis). **(D)** Location of the LG4 in the mRNA (x-axis) had an influence on the density of G triplets (y-axis). **(E)** Location of the LG4 in the mRNA (x-axis) is graphed to the density of G-quadruplets (y-axis). **(F)** The 5' LG4 mRNA (x-axis) density of G-quintuplets (y-axis) was compared to UTR, mid, and 3' introns density of G-quintuplets(unpaired t-test).

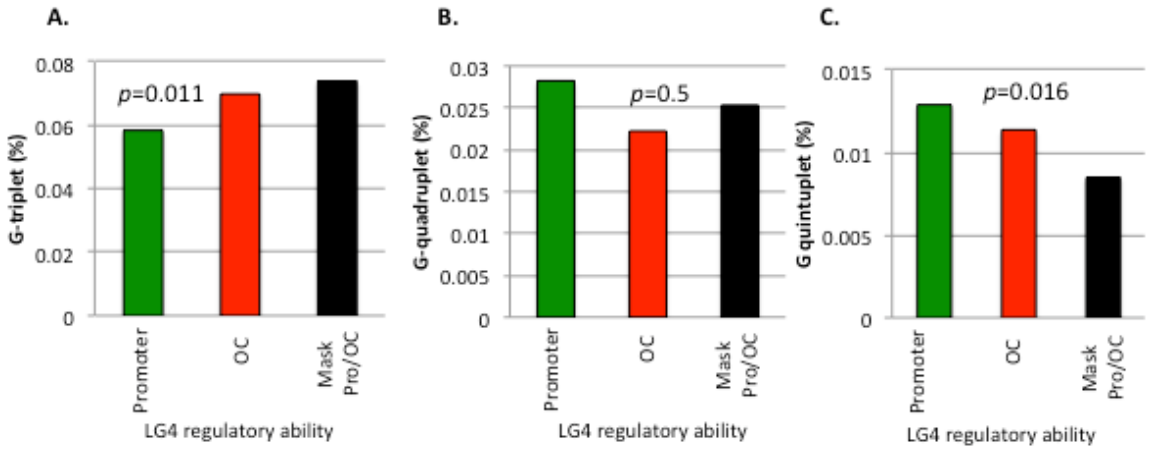

Figure 7. Correlation of Sequence with Length and Regulatory Ability. Outputs of Perl programs were used to identify relationships between regulatory ability of LG4s and the guanine repeat composition of the sequences. **(A)** The percentage of guanine triplets (G-triplets (%)) (y-axis) was graphed according to the LG4s corresponding regulatory ability (promoter, open chromatin (OC), or neither (mask Pro/OC) (x-axis). **(B)** In a similar fashion, the percentage of guanine quadruplets (G-quadruplets (%)) (y-axis) was graphed according to the LG4s corresponding regulatory ability (x-axis). **(C)** The percentage of guanine triplets (G-quintuplets (%)) (y-axis) was graphed according to the LG4s corresponding regulatory ability (x-axis).

Table 8. Oligonucleotides Tested by Circular Dichroism. LG4 sequences used to design oligonucleotide are listed (LG4 name) and corresponding sequence of oligonucleotide assayed (oligo sequence). The different G repeats involved in intra-molecular G4 are highlighted in different colors for each corresponding G-repeat length.

Table 8. Oligonucleotide tested by circular dichroism	
LG4 name	oligo sequence
MLF1	AGGGTGTGAGGTGAGGGGAAGGGTTGAGGGCGTGAGGTGAGGGG
F7	GTGGGGGATGGGGTGTGTGGGGTGTCGGGGAT
H2CN	CAGGGGTGGGGCTATGGGAGATCTGGGTGGGGTCAAGGCCCCAGGGATGGGGCT
TPRX1	CAGGGAGTGGGCCTGGGATCTGGGCTGGGCCTGGGATTGGGGCAGGGA
P2RX5	CTGGGCGGGGGGCGGGGGACGTTGGGAGGGTCCCTGGGGGCCCCGCT
PDE6B	GTGGGGTAGTGGGAGCAGCATGGGGTAGGGGAATAGCGTGGGGTA
SARDH	GAGGGAGGGTGAGGGAGAGGGAGGGTGAGGGAGAGGGAGGGTGAGGGGA
TMPRSS2	GAGGGGGCGAGGGGGTGAGTGAGGGGGCGAGGGGGTG
NACA	AGCTGGGGGAGTGGGGGCCCCTTTGGGGGGTGGGGTA
CTCFL	CAGGGGGGAAGGGGAGTAGGGAGGGGGAGGGGGAG
UBE2I	GAGGGAGGGAGGGAATGAGGGAGGGAGGGAAT
MXI1	GAGGGGAGGGGCGTGTGAGGGGAGGGGAG
TCF3	AGGGGGTGAGGCGGGAAGGGGACAGCAGAACTCACGGGGT
PRDM16	GAGGGGTGTGTTGGGAGGGGTGAGTTGGGGGGGTGCACCGGGAGGGGTG
TP73	CAGGGCTGGGATTGCGGGAGGAGGGGC
G=triplet G=quadruplet G=quintuplet G=sextuplet	

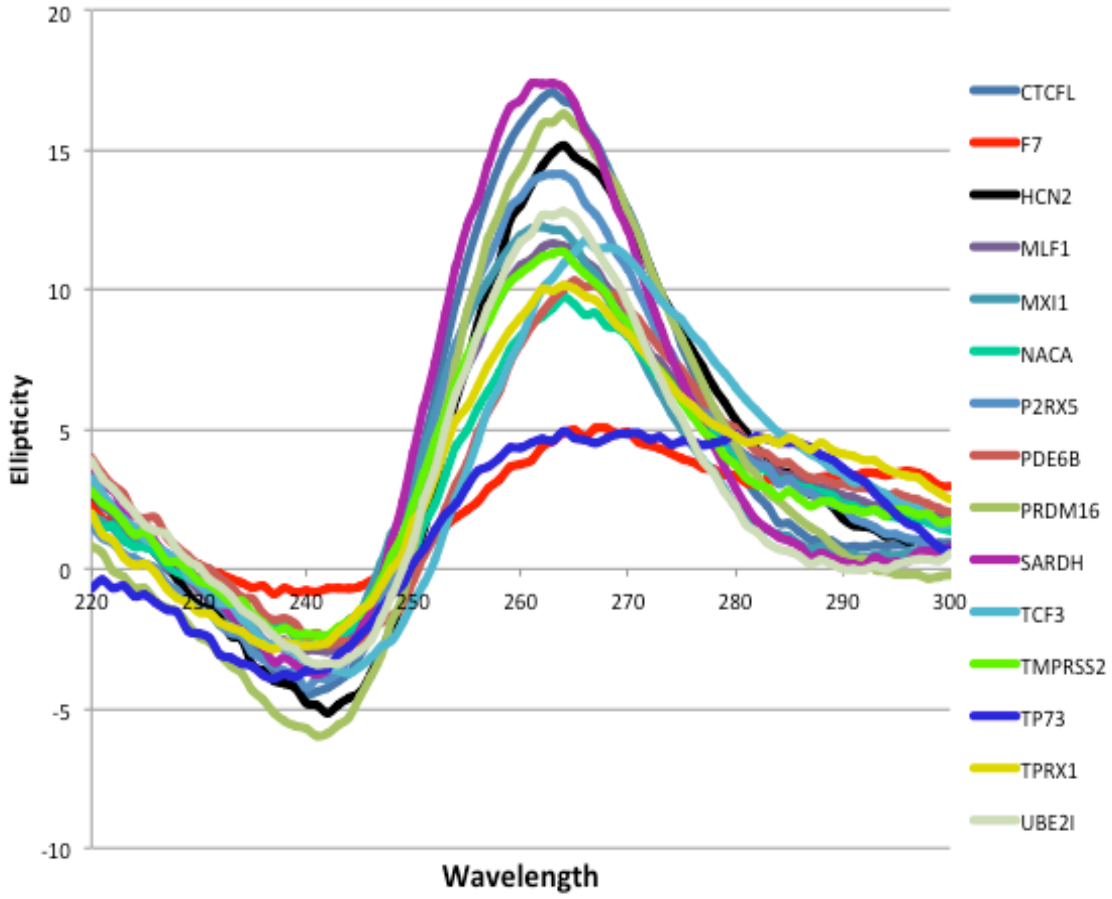

Figure 8. Circular Dichroism Ellipticities of Oligonucleotides Representing LG4s Display Spectra Consistent with G4 Formation. Each individual LG4 CD scan (Ellipticity) (y-axis) is displayed from a 220-300 wavelength (nm). Each LG4 assayed regardless of its guanine composition had characteristics of parallel G4, with a peak at ~260 nm and dip at ~240 nm. F7 (red line) was the only oligonucleotide assayed that displayed a CD spectra suggestive of antiparallel G4 formation, with a distinguishable peak at 295nm.

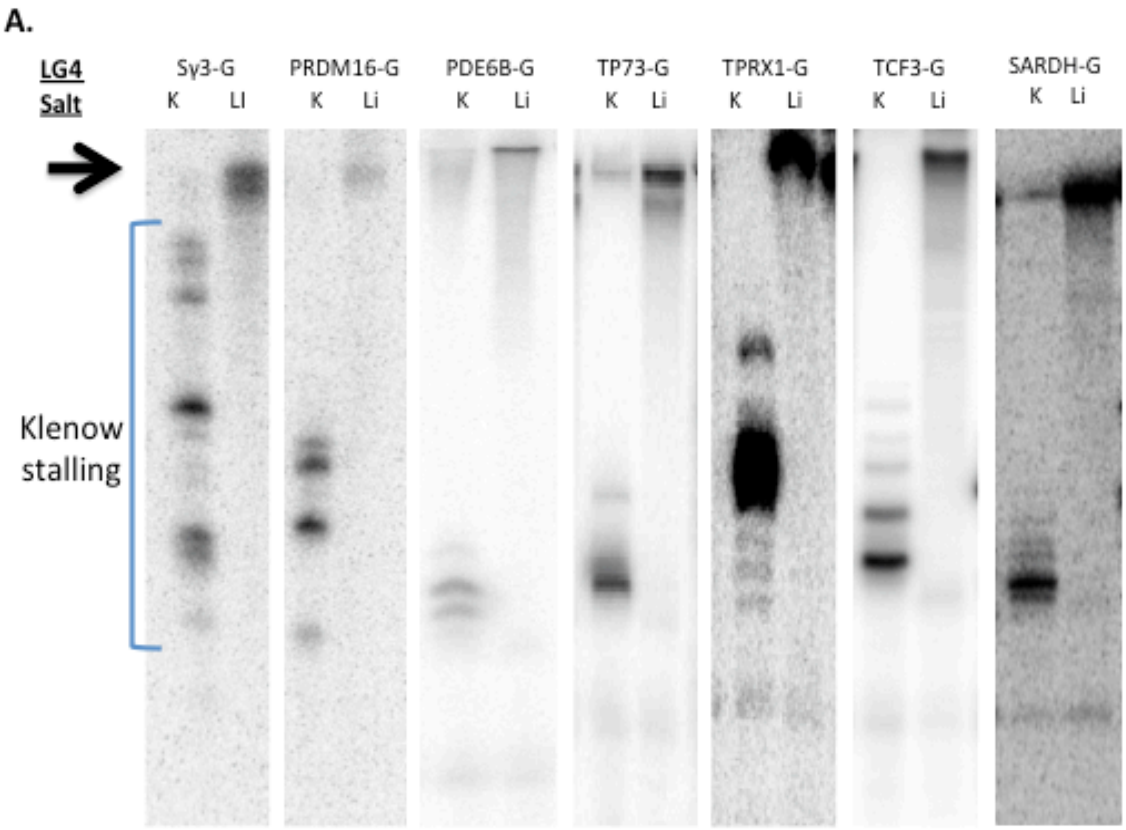

Figure 9. Klenow Primer Extension Reactions in K^+. The location of stalled Klenow is denoted by the bracket on the left side of the gel (Klenow Stalling). Completed replication products are denoted by the arrow. Reactions took place in G4 permissive conditions (K^+) or G4 non-permissive conditions (Li^+) **(A)** Subset of LG4 clone extension reactions displayed a K^+ dependent stalling on the G-rich strand, indicating G4 formation.

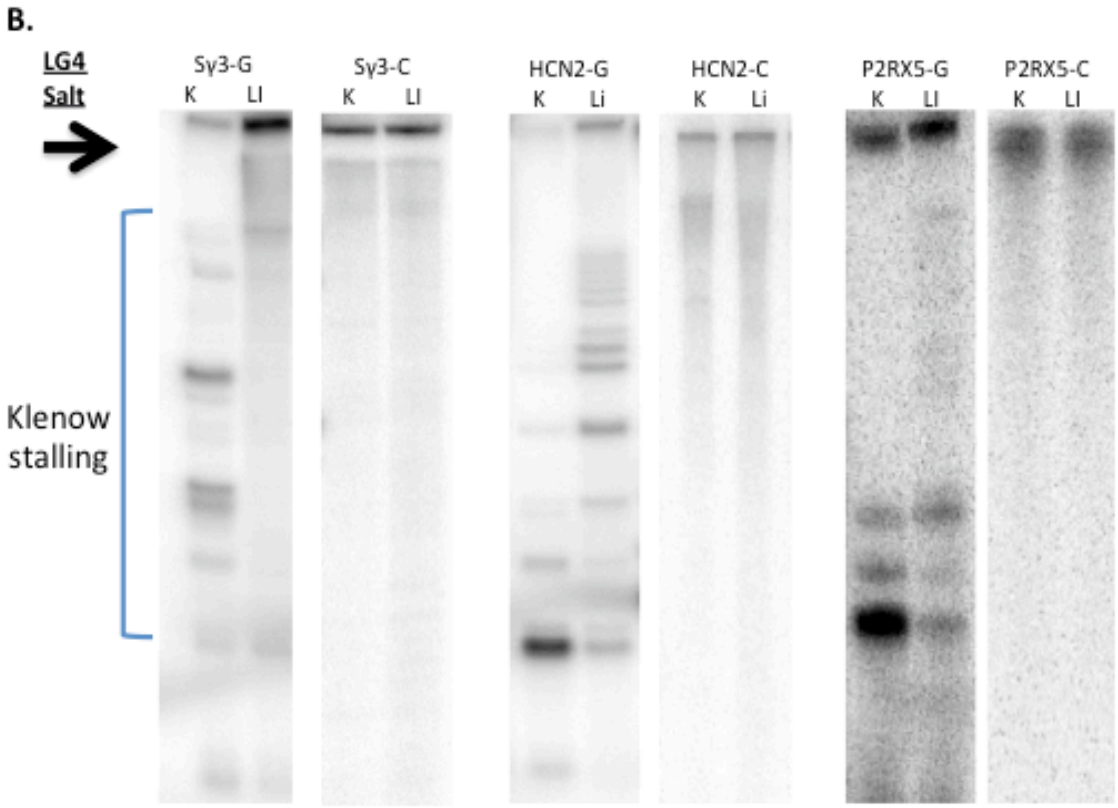

Figure 9. Klenow Primer Extension Reactions in Li^+. The location of stalled Klenow is denoted by the bracket on the left side of the gel (Klenow Stalling). Completed replication products are denoted by the arrow. Reactions took place in G4 permissive conditions (K^+) or G4 non-permissive conditions (Li^+). (B) Subset of LG4 clone extension reactions displayed a K^+ independent stalling on the G-rich strand (*HCN2*-G, *P2RX5*-G). However, did not stall in the C-rich strand (*HCN2*-C, *P2RX5*-C).

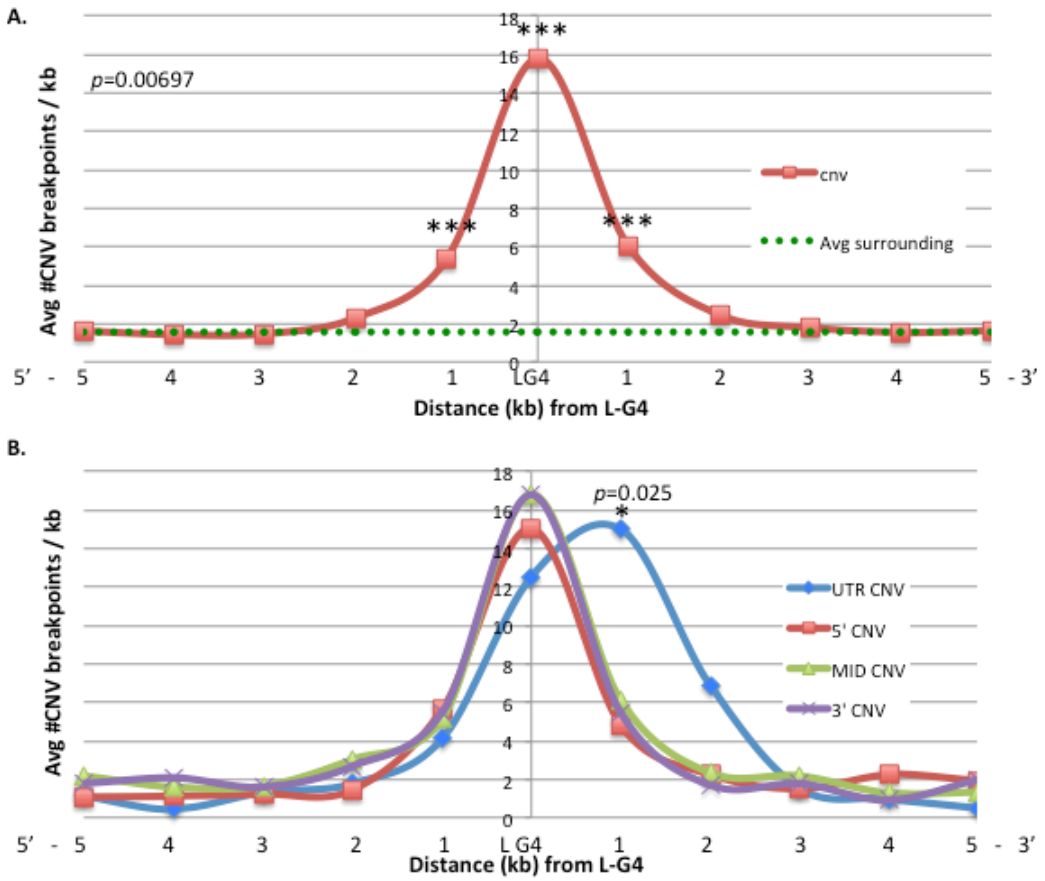

Figure 10A. CNV Breakpoint Densities. CNVs breakpoints were collected from dbVAR on NCBI.org and the average number of breakpoints/kb (y-axis) was calculated for each transcribed LG4 (LG4 central X axis), 1 kb increments both 5kb 5' and 3' (1-5 x-axis), and the rest of the transcript not directly associated with LG4 (average surrounding green dotted line). These are the precise increments used in Figure 4 **(A)** The average # of CNV breakpoints/ kb (CNV density) for all LG4s compared to surrounding regions is graphed (green line and red line 2-5 versus red line central x-axis). Outside of LG4s, there was an increase in CNV density in the 1 kb directly flanking both sides of LG4s (red line 5'-1 and 3'-1). **(B)** CNV density was graphed according to LG4 location in the mRNA. There was no significant variation between LG4 location and CNV density (LG4 central X axis). LG4s located in the UTR (blue line) had a significant increase of CNVs 1 kb directly 3' (1-3' x-axis)

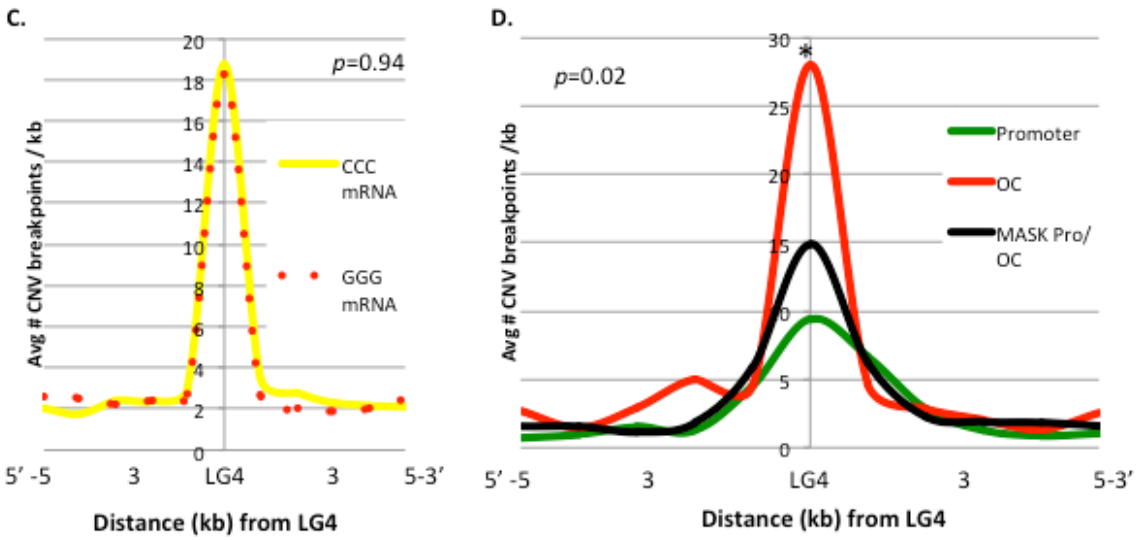

Figure 10B. CNV Breakpoint Densities. CNVs breakpoints were collected from dbVAR on NCBI.org and the average number of breakpoints/kb (y-axis) was calculated for each transcribed LG4 (LG4 central X axis), 1 kb increments both 5kb 5' and 3' (1-5 x-axis), and the rest of the transcript not directly associated with LG4 (average surrounding green dotted line). These are the precise increments used in Figure 4 **(C)** Sequences that support G4 are on the transcribed strand (central x-axis yellow line CCC mRNA), or non-transcribed strand (central x-axis red dotted line GGG-mRNA) and their density of CNV breakpoints/kb (y-axis) was graphed accordingly. **(D)** The average number of CNV breakpoints /kb (y-axis) in open chromatin LG4s (OC- red line central x-axis), promoter LG4s (Promoter green line LG4 central x-axis) and all other LG4 regions (Mask Pro/OC, black line LG4 central x-axis) is graphed.

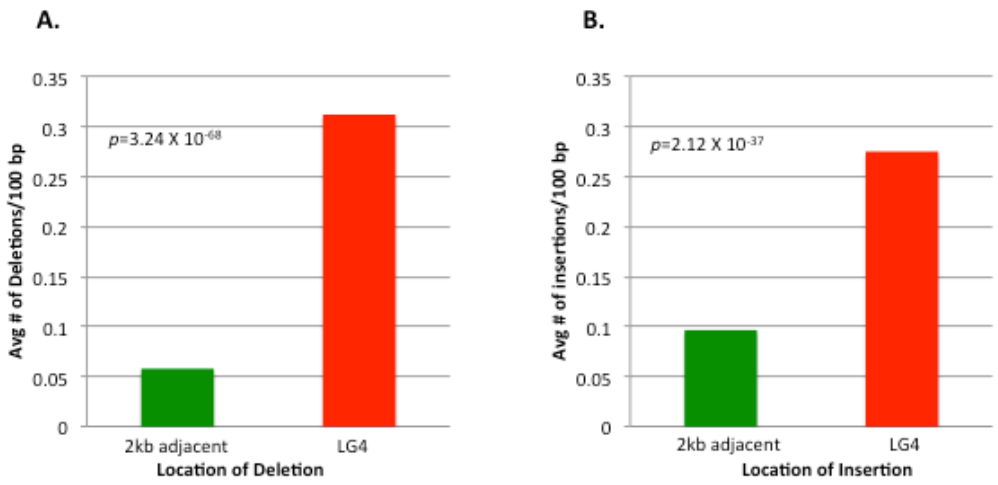

Figure 11. Indel Density in Transcribed Regions of LG4s. Entries from the dbSNP database were graphed according to individual quantity of insertion or deletion events/100 bp (y-axis). **(A)** The average number of deletions (y-axis) located within LG4s, or in the surrounding 2kb both 5' and 3' (x-axis) is graphed. **(B)** The average number of insertions located (y-axis) located within LG4s, or in the surrounding 2kb both 5' and 3' (x-axis) is graphed.

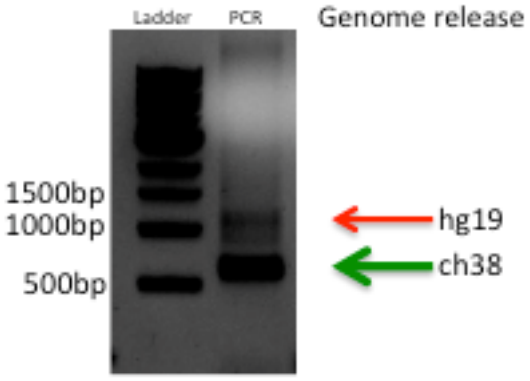

Figure 12. Human Genomic PCR Products of *CRLF2* LG4 Intron. PCR amplification of *CRLF2*'s LG4 from human, disease-free genomic DNA. Multiple sized products correspond to recent human genome release (ch38 green arrow), and the previous human genome release (hg19 red arrow).

Table 9. Transcribed LG4 Involved in Cancer.
List of proteins containing LG4s in their mRNA that are associated with the causation or progression of cancer. The name of the LG4 containing protein and the corresponding type of cancer(s) are shown.

LG4 Protein	Type of cancer
ABR	Medulloblastoma
ANO9	colrectal, lung, breast
CACNA2D2	Epilepsy, lung
CELSR3	marfan syndrome, gastric
CTCFL	prostate
CRLF2	Leukemia associated with downs syndrome
DPEP1	colrectal
FOXP1	lympohoma, breast , speech disorders
GCNT3	colrectal
IL3RA	hematologic malignancies, myloid leukemia
MBOAT7	bladder and breast
MLF1	Leukemia
MUC12	colrectal
MXI1	multiple sclerosis, Neurofibrosarcoma, Prostate
NKD2	soft tissue carcinoma, gastric
P2RX5	basal and squamous cell carcinomas, leukemia
PIWIL1	seminoma, gastric
PKP3	squamous cell , Lung ,
PRAME	neuroblastoma, leukemia
PTPRH	Peutz–Jeghers syndrome
SLC12A7	breast , Renal tubular acidosis
TCF3	leukemia
TMPRSS2	prostate
TP73	breast, cervical, colorectal, esophageal, stomach, endometrial
TUSC5	lung
UBE2I	ovarian
PRDM16	Leukemia, heart disease

Table 10. Transcribed LG4 Involved in Developmental Diseases.
List of proteins containing LG4s that are associated with the causation or progression of developmental disease. The name of the LG4 containing protein and the corresponding type of developmental disease(s) are shown.

LG4 protein	Type of developmental disease
LMF1	Lipase deficiency
NFATC1	heart anomalies
NOS3	Heart defects, cardiovascular disease, Alzheimer's
ROR2	Brachydactyly, type B1,Robinow syndrome
TNNI3	Cardiac hypertrophy,cardiomyopathy,left ventricular wall thickness
SARDH	Sarcosinemia
UROC1	urocanic aciduria
HBM	respiratory distress syndrome
PDE6B	retinitis pigmentosa
F7	cerebral palsy, sickle cell anemia, coagulation disorder
COL5A1	Ehlers-Danlos syndrome, Achilles tendon pathology
FAM20C	Raine syndrome
CANT1	Desbuquois dysplasia-1
CLDN14	deafness
BFSP2	Juvenile Cataracts
AHRR	azoospermia infertility, male oligospermia, endometriosis, micropenis

Table 11. Transcribed LG4 Involved in Neurological Diseases.
List of proteins containing LG4s that are associated with the causation or progression of neurological diseases. The name of the LG4 containing protein and the corresponding type of neurological disease(s) are shown.

LG4 Protein	Type of neurological disease
GRIN1	bipolar disorder, schizophrenia, seizures
HCN2	seizures, febrile
NCOR2	bipolar disorder,osteoarthritis,
OPRL1	Alzheimer disease, Coronary spasms, Placental abruption
PRKCB	autism, diabetes, nephropathy
RASA3	neurofibromatosis, and cerebritis
WBSCR17	Williams syndrome
ADCK3	Coenzyme Q10 deficiency,Spinocerebellar ataxia, autosomal recessive 9
ARHGEF10	Slowed nerve conduction velocity
CALY	attention deficit hyperactivity disorder
CPLX1	schizophrenia
CTDP1	CCFDN Congenital cataracts, facial dysmorphism, neuropathy
D2HGDH	D-2-hydroxyglutaric aciduria
DLGAP2	eplepsy, mental retardation
DVL1	Alzheimer's Disease
GALNS	Mucopolysaccharidosis IVA
MAP2K2	Cardiofaciocutaneous syndrome
NACA	alzheimers, down syndrome

CHAPTER III

FORMATION OF G-QUADRUPLEX DNA INFLUENCES THE GENETIC STABILITY OF HUMAN *TCF3* (*E2A*)

Abstract

The formation of highly stable four-stranded DNA, called G-quadruplex (G4), promotes site-specific genetic instability. Increasing experimental evidence connects G4 sequence motifs with specific gene rearrangements. The human *TCF3* gene (also termed *E2A*) is subject to genetic instability associated with severe disease, most notably a common translocation event t(1;19) associated with acute lymphoblastic leukemia. The sites of instability in *TCF3* are not randomly distributed, but focused to certain sequences. We asked if G4 DNA formation could explain why *TCF3* is prone to recombination and mutagenesis. Here we demonstrate that sequences surrounding the major t(1;19) break site and a region associated with copy number variations, both contain G4 sequence motifs. The motifs identified readily adopt G4 DNA structures that are stable enough to interfere with DNA synthesis under physiological conditions *in vitro*. When introduced into the yeast genome, *TCF3* G4 motifs promoted gross chromosomal rearrangements in a transcription-dependent manner. Our results provide a molecular rationale for the site-specific instability of human *TCF3,* suggesting that G4 DNA structures contribute to oncogenic DNA breaks and recombination.

Introduction

Transcription Factor 3 (TCF3), also called E2A, is a key regulatory protein that institutes transcriptional programs for proper B and T cell differentiation (Miyazaki et al., 2014; Kee et al., 2009). Given well-established regulatory roles in transcription activation, it is not surprising that disruption of *TCF3* or its gene product is associated with malignant transformation. For instance, a translocation event t(1;19) between the *TCF3* and *PBX1* (Pre-B cell leukemia homeoboX 1) genes results in the expression of a TCF3-PBX1 chimera, which is commonly found in acute lymphoblastic leukemia (ALL) (Hunger, 1996; Aspland et al., 2001; Pui et al., 2004). Genomic studies of Burkitt's lymphomas found *TCF3* to be among the most mutated genes (Schmitz et al., 2012) and 70% of patient samples were identified with heterozygous *TCF3* deletions in Sezary syndrome cells, an aggressive T cell lymphoma (Steininger et al., 2011). In addition to pre-B cell cancers, TCF3-PBX1 fusion transcripts have been identified in non-small cell lung cancer (Mo et al., 2013), indicating that the impact of this rearrangement likely extends beyond the immune system. While *TCF3* appears to be a hot spot for DNA breaks, and broadly associated with oncogenesis, the mechanisms responsible for driving this genetic instability are undefined.

The distribution of chromosomal breakpoints involved in the t(1;19) translocation imply that certain sequences are involved in the promoting instability. Previous characterization of t(1:19) breakpoints revealed that most of the recombination junctions (16 of 24) cluster in a 5 bp sequence window in *TCF3* (Wiemels et al., 2002). Those breaks were associated with CpG

sequences in *TCF3*, but not *PBX1* (Tsai et al., 2008), and these same recombination sites are surrounded by transposable element repeats, specifically a MER20 transposon (Rodić et al., 2013). The non-random distribution of break sites and proximity to DNA repeat motifs imply that the recombination events involving *TCF3* are influenced at either the DNA sequence or structural level.

DNA repeats can promote instability because of their ability to adopt non-B form structural conformations that interfere with normal DNA transactions (van Kregten and Tijsterman, 2014; Zhao et al., 2010; Lobachev et al., 2007). In particular, guanine repeats readily fold into highly stable four-strand structures called G-quadruplex (G4 DNA), which promote mutagenesis and recombination (Tarsounas and Tijsterman, 2013; Bochman et al., 2012; van Kregten and Tijsterman, 2014; Wu and Brosh 2010, Brooks et. al., 2010). G4 DNA structures are stabilized by hydrogen bonding between four guanine bases to create a single “tetrad” of guanines. When tandem guanine repeats are present, stacks of tetrads can form within or among DNA strands to build the four-stranded, or quadruplex, structure. The size, stability, and specific type of G4 structure depends upon the characteristics of the repeat sequence and aqueous conditions (Burge et al., 2006).

G4 DNA structures have been identified with loci involved in both induced and spontaneous genome instability (Maizels and Gray, 2013). At the guanine-rich immunoglobulin switch regions, programmed recombination requires transcriptional activation, and transcribed switch regions can form loops that are stabilized by RNA/DNA hybrids on one strand and G4 DNA on the other

(Duquette et al., 2004). Factors involved in this recombination pathway specifically recognize G4 DNA (Larson et al., 2005). In addition to induced recombination events, G4 structures lead to instability at other guanine-rich genomic loci. In a few examples, G4 structure formation was recently attributed to the instability of *HOX11* gene, and is associated with t(10;14) translocation breakpoints (Nambiar et al; 2013). Similarly, the break sites in *BCL2* leading to the t(14;18) translocation corresponded with G4 structure formation (Nambiar et al., 2011). This is not simply coincidental; experimental systems have directly connected instability of guanine-rich DNA with the formation of G4 structures. For instance, chromosomal rearrangements at guanine-rich human minisatellite sequences were dependent upon G4 formation (Piazza et al., 2012; Lopes et al., 2011), and addition of G4-stabilizing ligands increased the instability at the G4 repeats, but not for other types of sequence repeats (Piazza et al., 2010). In a similar vein, chromosomal rearrangement assays have been developed for use in characterizing specific loci and factors for G4-mediated genome instability (Yadav et al., 2014; Paeschke et al., 2013; Piazza et al., 2012).

Considering the emerging evidence connecting G4 DNA sequences with genome instability we asked if known rearrangements of the *TCF3* gene correlate with G4 structure formation. Here, we applied a comprehensive computational analysis of *TCF3* and the translocation partner *PBX1* and found that G4 motifs accompany sequences near the major t(1;19) breakpoints. Using multiple methods, we demonstrate here that those sequences fold into highly stable G4 structures *in vitro* and induced site-specific instability *in vivo*. We also

characterize a second site in *TCF3* for G4 formation that is not involved with the t(1;19) translocation, but is instead associated with copy number variations. Our results indicate that the site-specific instability of human *TCF3* is governed at least in part by a capacity to form structures at those sites.

Materials and Methods

Sequence Analysis

The *TCF3*/*PBX1* t(1:19) breakpoint sequences from the Lieber database (Wiemels et al., 2002; Tsai et al., 2008) and the Translocations in Cancer database (TICdb) (Novo et al., 2007) were mapped onto *TCF3* and *PBX1* genome sequences. Copy number variations (CNVs) for *TCF3* were downloaded from the database of genomic structural variation (dbVAR) on NCBI.org (Lappalainen et al., 2013). All CNVs' breakpoints (> 99 bp) were mapped to *TCF3*'s genomic location and confirmed using Ensembl release 77 (Kersey et al., 2014). For the identification of G4 sequence motifs, *TCF3* and *PBX1* sequences were analyzed using QGRS mapper (Kikin et al., 2006) with the following filters: A max loop length of 45 nucleotides, minimum G group of 3, and a loop size 0-36 nucleotides. The output of that analysis was mapped to *TCF3* and *PBX1* genes. G4 motifs with a QGRS score of at least 42 were presented as the number of independent G4 sequences identified in 2 kb non-overlapping windows. G4 repeat motifs used in structure analysis were selected based on their proximity to the t(1;19) breakpoint clusters in *TCF3* and *PBX1* (T-5', T-3', T-3'(2), P-1, and P-2), and to breaks associated with CNVs identified for *TCF3* (T-lg). The location of

all dbSNP database insertions and deletions (Sherry et al., 2001) were mapped to *TCF3* using Enseml release 77 (Kersey et al., 2014) and density was calculated by number of insertion or deletion events per 2 kb.

G4 Folding and PAGE Analysis

G4 oligonucleotides for native PAGE and CD analysis were synthesized and PAGE purified by Operon (Huntsville, AL). Sequences are shown in Table 12 and Table 13. For Native PAGE, oligonucleotides were 5' end labeled using T4 PNK (New England Biolabs, Ipswich, MA) and [γ-P^{32}]ATP (MP Biomedicals, Solon, OH) at 37 °C for 30 minutes. Unincorporated label was removed by Illustra Microspin G-25 spin chromatography (GE healthcare, Pittsburgh, PA). G4 structures were formed in reactions containing 100 mM KCL in Tris-EDTA buffer, that were initially denatured by incubating in a small > 90°C water bath and then allowed to cool in the bath slowly to room temperature. Samples were then incubated an additional hour at 37°C. Native PAGE experiments used 16% polyacrylamide (37:1) containing 0.5 X TBE with 100 mM KCL in the gel and run buffer. Oligonucleotides were resolved by electrophoresis at 100 V at room temperature for 6 hours. Denaturing PAGE of radiolabeled oligonucleotides used 16% polyacrylamide (19:1) gels made with 7 M urea and 0.5X TBE. Prior to loading, samples were denatured in 90% formamide and heated to 90˚C for 20 minutes. DNA was resolved by electrophoresis at 400V for 1.5 hours. Images were captured by phosphorimaging using a Molecular Dynamics Storm 840 phosphorimager (Amersham/GE).

Circular Dichroism

CD analysis was performed using an Aviv model 215 CD spectrometer at 37°C. Spectra were taken in 1 cm path quartz cells containing 12 μM G4 or GT oligonucleotide in 10 mM Tris–HCl, pH 7.6, 1 mM EDTA, and 100 mM KCl. The molar ellipticity was measured from 220–300 nm and recorded for 3 scans in 1 nm increments at a 1 second averaging time.

Primer Extension Assays

Phagemids for extension assays were obtained by reconstituting the genomic sequence via overlapping PCR using semi-complimentary primers in a standard PCR reaction. PCR products were gel purified and TOPO cloned (Invitrogen) into pCR2.1 in both orientations and verified by sequencing. Templates for extension assays included the G4 motifs shown in Table 12 and were the following sizes; T-5' (161 bp), T-3' (472 bp), T-3'(2) (168 bp), T-lg (124 bp), P-1 (95 bp), P-2 (92 bp) (Figure 18). Closed circular single-stranded DNA was obtained using M13K07 helper phage (NEB) according to the manufacturer's instructions.

Polymerase extension assays were performed essentially as described (Ehrat et al., 2012) and based on previous G4 assays (Sun and Hurley, 2010; Weitzmann et al., 1996). Single-stranded phagemid templates were primed with a ^{32}P 5' end labeled M13 forward primer, which was extended with Klenow or Taq polymerase (NEB). In Klenow reactions, KCl or LiCl was added to a final concentration of 25 mM. Klenow extension reactions took place at 37 °C for 8 minutes on single-stranded template primed with 5' end-labeled M13 forward (-

20) primer. Taq reactions used identical conditions, however temperature ranged from 50°C-80°C for 9 minutes in buffers containing either $(NH_4)_2SO_4$ or KCl salt. Extension reactions were stopped by the addition of an equal volume of 90% formamide and 1mM EDTA followed by heating to 90°C for 20 minutes. Products of polymerase extension were resolved by 8% denaturing PAGE (19:1) with 7 M urea and 0.5X TBE, at 700 V at room temperature. Gels were then dried and images were captured with a Molecular Dynamics Storm 840 phosphorimager (Amersham/GE).

Gross Chromosomal Rearrangements Assay

The plasmids containing *lys2 T-G-Top* and *lys2 T-G-Btm* cassettes were constructed by inserting the 472 bp T-3' sequence (Human genome 1:1618084-1618556) into the *Bgl*II site located +390 nucleotide position within the S. cerevisiae *LYS2* gene in two different orientations relative to transcription start site. Using the standard two-step allele replacement protocol, these constructs were used to replace the wildtype *LYS2* gene located proximal to *CAN1* on the left arm of the chromosome V. The rates of GCR occurring at the chromosome V were determined according to the previously described procedure (Yadav et al., 2014). Briefly, individual colonies were inoculated into the rich media (1% yeast extract, 2% petone and 2% glucose - YEPD) and cultured to saturation at 30°C**.** Appropriate dilution of the cultures was plated either on the YEPD media for determination of total cell numbers or on the selective media (synthetic complete media supplemented with canavanine (60 mg/l) and 5-Fluoroorotic acid (1 g/l))

for determination of the cells that lost *CAN1* and *URA3* genes. For each strain, 24 to 32 cultures were used to calculate rate and 95% confidence levels either by Lea-Coulson method of median (*top1*Δ*,* high transcription) (Spell and Jinks-Robertson, 2004) or by P_0 method (*wt*, high transcription, *top1*Δ*,* low transcription) (Foster, 2006).

Results and Discussion

G4 Sequence Motifs Surround Regions of Instability in *TCF3*

Based on a growing body of evidence linking G4 structures with site-specific genetic instability, we asked if formation of G4 DNA structures could explain the apparent genetic instability focused within the *TCF3* gene. The positions of 30 different breakpoints in *TCF3* spanning 4 kb, and 16 breakpoints in *PBX1* spanning 12 kb, are documented in the Translocations In Cancer database (TICdb) (Novo et al., 2007) and the Leiber database (Tsai et al., 2008). We used these sites to analyze and map break positions in relation to G4 sequence motifs. This was accomplished by applying a web-based server program for G4 structure prediction to overlay G4 sequences with known break site positions.

Quadruplex forming G-Rich Sequences (QGRS mapper) is a web-based server program developed to score nucleic acid sequences for G4 forming potential (Kikin et al., 2006). Single-stranded DNA can adopt stable G4 structures so long as continuous or patterns of tandem guanine repeats are present, and the density of repeats impacts the overall stability of the structure (Sen and Gilbert, 1988, 1990). QGRS mapper converts input sequences into scores

representing the likelihood for G4 formation, with a maximum score value of 180 given for a repeat of guanines that is 27 nucleotides long (Kikin et al., 2006). The location of G4 motifs within the queried sequence is an output of the program (Kikin et al., 2006). Figure 13A and 13B diagrams the output for QGRS scoring of *TCF3* and *PBX1* genes. Both strands were analyzed for non-overlapping motifs. Two regions of *TCF3* are associated with genetic instability; the t(1;19) translocation site (which forms the TCF3-PBX1 chimera), and a region rich in breaks connected to CNVs. Both of these regions are intronic. QGRS mapper identified 25 different G4 sequences at the CNV site, and 15 G4 sequences surrounding the major t(1;19) break site (Figure 13A). An additional region of high G4 sequence density can be found in the 5' end of the gene, but this site did not correlate with any known breakpoints, suggesting the genetic positioning of G4 motifs influences structure formation, or that instability at that site is not connected with diseases cataloged in existing databases. *PBX1* also contains G4 motifs near the t(1;19) translocation site, but G4 motifs occur at much lower density compared to *TCF3*, with 2 found for *PBX1* (Figure 13B) compared to 15 for *TCF3* (Figure 13A).

TCF3* and *PBX1* G4 Motifs Support G4 Structure Formation *In Vitro

Considering the breadth of experimental evidence linking site-specific instability with G4 DNA structures (Tarsounas and Tijsterman 2013; Bochman et al., 2012; van Kregten and Tijsterman 2014; Wu and Brosh 2010; Brooks et. al., 2010), we asked if the guanine-rich motifs we identified using QGRS mapper also form

stable G4 structures under physiological conditions. We selected multiple G4 motifs for each break site locus. This includes one repeating G4 motif from a large guanine-rich intron connected with CNVs in *TCF3* (T-lg), three motifs flanking the t(1;19) breakpoint (T-5', T-3' and T-3'(2)), and two motifs next to the *PBX1* t(1;19) breakpoints (P-1 and P-2). Sequence and QGRS scores for each motif we tested are listed in Table 12. Importantly, the sequence "T-5" is 20 bases 5' and "T-3" is just 2 bases 3' of a t(1;19) breakpoint cluster. T-3'(2) is a second G4 motif located ~ 1200 bp from the major break site cluster. P-1 and P-2 are 70 and 750 bp from a t(1;19) break site cluster, respectively. However, unlike the breakpoints in the *TCF3* locus, the breakpoints in intron 1 of *PBX1,* involved with *TCF3-PBX1* fusions, are more broadly distributed (Wiemels et al., 2002), suggesting that the G4 motifs residing within the *PBX1* locus may not instigate translocations with *TCF3* to same degree as the *TCF3* G4 sequences.

Each single-stranded guanine-rich motif (termed "G4") was synthesized along with a companion control in which tandem guanine repeats of three or more were disrupted by substituting thymine, thereby greatly reducing the potential for G4 formation (termed "GT") (Table 13). GT control and G4 motif oligonucleotides co-migrated on denaturing PAGE (Figure 14A top), as expected. G4 DNA is stabilized by K^+ or Na^+ ions (Sen and Gilbert 1990; Williamson et al., 1989). In the presence of 100 mM KCl, all of the *TCF3* and *PBX1* G4 oligonucleotides migrated as larger and smaller species compared to the GT interrupted control, which migrated as a single product in native PAGE (Figure 14A bottom). The slow migrating species are consistent with inter-molecular

structures, and species migrating faster than the GT control are consistent with intra-molecular, or self-pairing, conformations. GT control oligonucleotides retained identical mobility patterns in Native PAGE experiments independent of the presence of KCL (not shown), as expected. T-5' retains some self-complementarity even when the guanine repeats were disrupted by thymine (GT control), likely explaining the faster mobility pattern observed in Native PAGE. Consistent with an intra-molecular structural conformation, scrambling the T-5' sequence (T-5' S) reduced its mobility upon Native PAGE compared to the companion GT and G4 samples (Figure 14A, bottom left). Together, we conclude that the guanine-rich sequences derived from *TCF3* and *PBX1* break site regions adopt alternative DNA conformations in the presence of K^+. This is consistent with G4 DNA.

We further tested the ability of each *TCF3* and *PBX1* sequence motif to adopt G4 DNA in solution using circular dichroism (CD), comparing spectra of those oligonucleotides with controls that cannot adopt stable G4 conformations. CD measures the differential absorption of circularly polarized light by chiral molecules (called ellipticity). It is best suited for identifying the presence of structures, like G4, but not sensitive enough for atomic-level resolution. G4 DNA structures produce characteristic CD spectra, with positive peaks at either 264 or 295 nm and negative dips at 265 or 240 nm, respectively, depending on the type of quadruplex (Balagurumoorthy et al., 1992; Kypr et al., 2009; Đapić et al., 2003; Vorlickova et al., 2012). The CD spectra for all *TCF3* motifs tested showed molar ellipticities that peak at ~264 nm and a dip at ~240 nm (Figure 14B), consistent

with G4 DNA. Interestingly, T-5' shows a shallow and broadened peak that extends beyond 280 nm (Figure 14B, top left), probably reflecting the presence of non-G4, or B-form variant, structures (Kypr et al., 2009), which is consistent with Native PAGE analysis for this oligonucleotide (Figure 14A). CD analysis of T-5' at a two-fold higher concentration resulted in larger 260 nm maximum and 240 minimum peaks (Figure 17), suggesting that at the lower concentration B-form conformations (i.e., self pairing) either reduces the potential for or competes with G4 DNA structure formation. Either way, the T-5' sequence appears to adopt multiple DNA conformations that deviate from standard duplex. *PBX1* sequences P-1 and P-2 both showed CD spectra comparable to *TCF3* G4 DNA (Figure 14B, top right). As expected, interruption of the tandem guanine repeats by thymine substitution (Table 13) eliminated the characteristic CD spectra for G4 (Figure 14B, bottom). We conclude that the sequences surrounding major break site regions in *TCF3* and *PBX1* adopt G4 DNA conformations in solution.

TCF3 and *PBX1* G4 Structures Block DNA Synthesis *In Vitro*

The precise mechanisms by which formation of G4 DNA induces genome instability are not defined. However, regions of repetitive DNA that become transiently denatured have an opportunity to interact, or self pair, to form structures that interfere with DNA metabolism (Lopes et al., 2011; Zhao et al., 2010; van Kregten and Tijsterman, 2014). Presumably, highly stable non-B form DNA structures are more likely to interfere with DNA transactions compared with those that are less stable. To test that model for *TCF3* G4 motifs we next

employed primer extension assays using genomic sequences surrounding the major break sites. This assay has been well described for characterizing G4 formation, showing that K^+ ions support guanine-dependent G4 formation and polymerase pausing, while Li^+ and NH_4^+ ions do not (Weitzmann et al., 1996; Sun and Hurley, 2010; Ehrat et al., 2012). Using that system, we expected to find K^+-dependent polymerase pausing on templates containing *TCF3* and *PBX1* break site sequences. Templates for this assay contained the G4 motifs shown in Table 12, but also some additional genomic sequence (Figure 18). We cloned each sequence into the plasmid pCR2.1 (Invitrogen, Carlsbad, CA) in both orientations with respect to an F1 origin and the primer-binding site. Single-stranded phagemids were then isolated for each orientation (cytosine-rich or guanine-rich sequences) and used as single-stranded templates for primer extension reactions catalyzed by Klenow polymerase. Independent of specific template, polymerase extension reactions showed full-length product when the cytosine-rich strands (complements to the G4 motifs) were assayed, and synthesis was independent of the salt used, as expected (Figure 15A, left and Figure 19). In contrast, extension reactions prematurely paused on the guanine-rich templates when K^+, but not Li^+, was present (Figure 15A, right). Although fully consistent with G4 DNA, we further tested the dependence of stalling on guanine by using thymine substitution mutagenesis to disrupt the repeats in the T-5', P-1 and P-2 templates (Figure 18). Full extension was restored when the G4 motifs were disrupted (Figure 20), essentially matching results for the C-rich complements (Figure 15A), and demonstrating that the guanine repeats are

needed for the polymerase pausing. Therefore, this stalling of synthesis argues that G4 structures can form from the *TCF3* and *PBX1* break site sequences.

We next asked if the G4 structures formed within *TCF3* and *PBX1* sequences are thermally stable, and thereby capable of adopting difficult to resolve structural conformations in the cell. Taq polymerase, used in standard PCR, has optimal activity around 75˚C (Lawyer et al., 1993). One would predict that simple hairpins would denature at elevated temperatures, and beyond 37 ˚C any polymerase pausing that does occur would reflect the presence of highly stable template blockades. Based on that logic, we replicated the primer extension experiments described above with Taq at ranging temperatures in reactions containing K^+ salt (permissive for G4) or NH_4^+ salt (G4 disrupting). Resolution of the Taq extension products by denaturing PAGE revealed polymerase-stalling patterns similar to that of Klenow (Figure 15B). Importantly, bands corresponding to stalled synthesis on the guanine-rich templates were only marginally altered when reaction temperatures reached 80˚C, suggesting that the replication blockades formed within the *TCF3* and *PBX1* templates are thermally stable (Figure 15B). Full extension was observed when NH_4^+ was substituted for K^+ (Figure 15B) or when the C-rich strand served as the template (Figure 19), as expected. We conclude that the G-rich sequence motifs located proximal to the *TCF3* and *PBX1* t(1;19) translocation sites (Figure 13) and to the *TCF3* CNV break sites (Figure 13A) all form G4 structures capable of interfering with DNA synthesis *in vitro*. A sensible model from these results, given the growing body of evidence that G4 structures promote recombination (Wang et

al., 2004; Koole et al., 2014; Yadav et al., 2014; Zhao et al., 2010; Katapadi et al., 2012; Nambiar et al., 2011, 2013) is that G4 structures in *TCF3* and *PBX1* promote site-specific translocation events and mutagenesis.

TCF3 Break Site G4 Motifs Induce DNA Breaks *In Vivo*

Although our *in vitro* results suggests that sequences proximal to *TCF3* breaks adopt G4 DNA, it does not necessarily connect structure formation with site-specific instability in the cell. Therefore, we next asked if the G4 motifs surrounding the *TCF3* t(1;19) break site promote instability *in vivo* using a previously described yeast genetic assay. In these systems, a model G4-forming sequence from the murine Sμ Ig switch region was shown to enhance ectopic recombination (Kim and Jinks-Robertson, 2011) and gross chromosome rearrangements (GCRs) (Yadav et al., 2014). The genome instability occurring at the Sμ Ig switch region was influenced by conditions that impact G4 structure formation, such as high transcription rate, orientation of the sequence with respect to the promoter, and disruption of structure metabolizing enzymes like Sgs1, a G4 specific helicase (Huber et al., 2002), or Top1 topoisomerase (Yadav et al., 2014). In order to test whether G4 motifs identified at the *TCF3* translocation breakpoints can also induce genome instability, we modified the GCR assay, which selects for the simultaneous loss of the *CAN1* and *URA3* genes located telomeric to a reporter cassette that is integrated on Chromosome V. This cassette consists of *LYS2* gene transcribed from the tetracycline-regulatable promoter *pTET*. A 472 bp sequence (T-3'-G4 Figure 18) surrounding

the *TCF3* t(1;19) G4 motif was introduced into this cassette so that the G-rich strand is positioned on either the non-transcribed (T-G-Top) or transcribed (T-G-Btm) strand with respect to the *pTET* promoter. In wildtype backgrounds under high transcription conditions (WT), the rates of GCR for T-G-Top and T-G-Btm did not differ significantly (Figure 16A). GCR rates increased significantly for both T-G-Top and T-G-Btm by 69- and 36-fold, respectively, in Top1-deficient yeast strains. This is similar to the G4-associated genetic instability observed for the model G4 sequence, Ig Sμ (Yadav et al., 2014). When the transcription from *pTET* was repressed by addition of tetracycline analog doxycycline (+DX), the rates of GCR for T-G-Top and T-G-Btm were both reduced by ~4-fold compared to the high transcription conditions (Figure 16A). The difference in instability between T-G-Top and T-G-Btm in both transcription conditions ranged from ~2.7-3 fold. Therefore, in this assay T-G-Top, which is in a G4 favorable orientation, promoted the formation of DNA breaks leading to chromosomal rearrangements. However, transcriptional orientation had less of an impact on instability compared to the model Sμ G4 sequence, with about a 28 fold difference between the two Sμ transcriptional orientations (Yadav et al., 2014) compared to 3 fold difference for *TCF3* under high transcription conditions (Figure 16A). This probably reflects the difference in the density of G-repeats and the relative sizes of the *TCF3* and Sμ sequences used in the assay. Regardless, the t(1;19) G4 break site sequence from *TCF3* displayed co-transcriptional genetic instability, fully consistent with the model that G4 DNA formation promotes *TCF3* instability.

The second site of instability in *TCF3* (T-lg) is associated with CNVs (Figure 13), and those sequences readily adopted G4 structures *in vitro* (Figure 14 and 15). We next asked if this site in *TCF3* contains signatures for G4-mediated instability by using the existing genome variation databases. T-lg is extensively repetitive (with 106 GGG repeats), so it is reasonable to predict that the CNVs mapping to this region are associated with DNA breaks related to the G4 motifs and G4 structure formation. If so, we would expect to find DNA break signatures proximal to the tandem guanine repeats. Therefore, we examined the entire *TCF3* gene using the human database for short genetic variation (dbSNP) (Sherry et al., 2001) available on Ensembl77 (Kersey et al., 2014), looking specifically to map the positions of short (<100 bp) insertions and deletions. At the CNV site (T-lg), there are approximately 8-fold more insertions and deletions (indels) compared to the rest of the gene's average (Figure 16B). The locations of the sequence variations are not random, with 93% of indels directly next to, or inside a tandem guanine repeat (Figure 21). Fine mapping of deletion positions shows that all directly flank or reside within the same guanine repeat sequence, which is part of a sequence motif repeated 32 times in the T-lg sequence (Figure 22). Insertion mutations were also distributed within 2 nucleotides of the tandem guanine repeats, but there are fewer insertions overall compared to deletions (Figure 22). Although it is not immediately clear what the pattern or type of sequence variation indicates with regard to the mechanism(s) of instability, the loss and gain of sequence coincides with repetitive guanines.

While it is possible that G4 motifs identified here promote instability simply due to the repetitive nature of the sequence, we favor a model whereby guanine repeats participate in alternate structure formation. Failure to resolve those structures promotes DNA breaks, which are manifested in the available genome databases as copy number variations and translocation events. This is significant because it is currently unclear why the *TCF3* gene is unstable. Our results provide a molecular rationale for the apparent instability of *TCF3*, suggesting that it is not the sequence of the break sites that explains the site-specific instability per se, but rather their capacity to adopt DNA conformations that are difficult for the cell to resolve. Our results build upon a growing body of evidence directly connecting G4 DNA with DNA breaks, suggesting a key role for the structure in promoting genetic instability, particularly at guanine-rich oncogenes.

References

Aspland, S. E., Bendall, H. H., and Murre, C. (2001). The role of E2A-*PBX1* in leukemogenesis. *Oncogene, 20*, 5708-5717.

Balagurumoorthy, P., Brahmachari, S. K., Mohanty, D., Bansal, M., and Sasisekharan, V. (1992). Hairpin and parallel quartet structures for telomeric sequences. *Nucleic Acids Research*, *20*(15), 4061-4067.

Bochman, M. L., Paeschke, K., and Zakian, V. A. (2012). DNA secondary structures: stability and function of G-quadruplex structures. *Nature Reviews Genetics*, *13*(11), 770-780.

Brooks, T. A., Kendrick, S., and Hurley, L. (2010). Making sense of G-quadruplex and i-motif functions in oncogene promoters. *Febs Journal*, *277*(17), 3459-3469.

Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K., and Neidle, S. (2006). Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Research*, *34*(19), 5402-5415.

Đapić, V., Abdomerović, V., Marrington, R., Peberdy, J., Rodger, A., Trent, J. O., & Bates, P. J. (2003). Biophysical and biological properties of quadruplex oligodeoxyribonucleotides. *Nucleic Acids Research*, *31*(8), 2097-2107.

Duquette, M. L., Handa, P., Vincent, J. A., Taylor, A. F., & Maizels, N. (2004). Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. *Genes & development*, 18(13), 1618-1629.

Ehrat, E. A., Johnson, B. R., Williams, J. D., Borchert, G. M., and Larson, E. D. (2012). G-quadruplex recognition activities of E. Coli MutS. *BMC molecular biology*, *13*(1), 23.

Foster, P.L. (2006). Methods for determinine spontaneous mutation rates. *Methods in Enzymology* 409:195-213.

Huber, M. D., Lee, D. C., and Maizels, N. (2002). G4 DNA unwinding by BLM and Sgs1p: substrate specificity and substrate-specific inhibition. *Nucleic Acids Research*, *30*(18), 3954-3961.

Hunger, S. P. (1996). Chromosomal translocations involving the E2A gene in acute lymphoblastic leukemia: clinical features and molecular pathogenesis. *Blood*, *87*(4), 1211-1224.

Katapadi, V. K., Nambiar, M., and Raghavan, S. C. (2012). Potential G-quadruplex formation at breakpoint regions of chromosomal translocations in cancer may explain their fragility. *Genomics*, *100*(2), 72-80.

Kee, B. L., Quong, M. W., and Murre, C. (2000). E2A proteins: Essential regulators at multiple stages of B-cell development. *Immunological reviews*, *175*(1), 138-149.

Kersey, P. J., Allen, J. E., Christensen, M., Davis, P., Falin, L. J., Grabmueller, C., and Staines, D. M. (2014). Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Research*, *42*(D1), D546-D552.

Kikin, O., D'Antonio, L., and Bagga, P. S. (2006). QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Research*, *34*(suppl 2), W676-W682.

Kim, N., & Jinks-Robertson, S. (2011). Guanine repeat-containing sequences confer transcription-dependent instability in an orientation-specific manner in yeast. *DNA repair*, *10*(9), 953-960.

Koole, W., van Schendel, R., Karambelas, A. E., van Heteren, J. T., Okihara, K. L., and Tijsterman, M. (2014). A Polymerase Theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites. *Nature communications*, *5*.

Kypr, J., Kejnovská, I., Renčiuk, D., and Vorlíčková, M. (2009). Circular dichroism and conformational polymorphism of DNA. *Nucleic Acids Research*, *37*(6), 1713-1725.

Larson, E. D., Duquette, M. L., Cummings, W. J., Streiff, R. J., & Maizels, N. (2005). MutSα binds to and promotes synapsis of transcriptionally activated immunoglobulin switch regions. *Current biology*, *15*(5), 470-474.

Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., and Church, D. M. (2013). DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Research*, *41*(D1), D936-D941.

Lobachev, K. S., Rattray, A., and Narayanan, V. (2007). Hairpin-and cruciform-mediated chromosome breakage: causes and consequences in eukaryotic cells. *Front Biosci*, *12*, 4208-4220.

Lopes, J., Piazza, A., Bermejo, R., Kriegsman, B., Colosio, A., Teulade-Fichou, M. P., and Nicolas, A. (2011). G-quadruplex-induced instability during leading-strand replication. *The EMBO journal*, *30*(19), 4033-4046.

Lawyer, F. C., Stoffel, S., Saiki, R. K., Chang, S. Y., Landre, P. A., Abramson, R. D., & Gelfand, D. H. (1993). High-level expression, purification, and enzymatic characterization of full-length Thermus aquaticus DNA polymerase and a truncated form deficient in 5'to 3'exonuclease activity. *Genome research*, *2*(4), 275-287.

Maizels, N., and Gray, L. T. (2013). The G4 genome. *PLoS genetics*, *9*(4), e1003468.

Mo, M. L., Chen, Z., Zhou, H. M., Li, H., Hirata, T., Jablons, D. M., and He, B. (2013). Detection of E2A-*PBX1* fusion transcripts in human non-small-cell lung cancer. *J Exp Clin Cancer Res*, *32*, 29.

Miyazaki, M., Miyazaki, K., Chen, S., Itoi, M., Miller, M., Lu, L. F., Varki, N., Chang, A., Broide, D.H., Murre, C. (2014). Copia autorizada por CDR. *Nature Immunology*, *15*(8), 767.

Nambiar, M., Goldsmith, G., Moorthy, B. T., Lieber, M. R., Joshi, M. V., Choudhary, B., and Raghavan, S. C. (2011). Formation of a G-quadruplex at the BCL2 major breakpoint region of the t (14; 18) translocation in follicular lymphoma. *Nucleic Acids Research*, *39*(3), 936-948.

Nambiar, M., Srivastava, M., Gopalakrishnan, V., Sankaran, S. K., and Raghavan, S. C. (2013). G-quadruplex structures formed at the HOX11 breakpoint region contribute to its fragility during t (10; 14) translocation in T-cell leukemia. *Molecular and cellular biology*, *33*(21), 4266-4281.

Novo, F. J., de Mendíbil, I. O., and Vizmanos, J. L. (2007). TICdb: a collection of gene-mapped translocation breakpoints in cancer. *BMC genomics*, *8*(1), 33.

Paeschke, K., Bochman, M. L., Garcia, P. D., Cejka, P., Friedman, K. L., Kowalczykowski, S. C., & Zakian, V. A. (2013). Pif1 family helicases suppress genome instability at G-quadruplex motifs. *Nature*, *497*(7450), 458-462.

Piazza, A., Boulé, J. B., Lopes, J., Mingo, K., Largy, E., Teulade-Fichou, M. P., and Nicolas, A. (2010). Genetic instability triggered by G-quadruplex interacting Phen-DC compounds in Saccharomyces cerevisiae. *Nucleic Acids Research*, *38*(13), 4337-4348.

Piazza, A., Serero, A., Boule, J. B., Legoix-Ne, P., Lopes, J., and Nicolas, A. (2012). Stimulation of gross chromosomal rearrangements by the human CEB1 and CEB25 minisatellites in Saccharomyces cerevisiae depends on G-quadruplexes or Cdc13. *PLoS genetics*, *8*(11), e1003033.

Pui, C. H., Relling, M. V., and Downing, J. R. (2004). Acute lymphoblastic leukemia. *New England Journal of Medicine*, *350*(15), 1535-1548.

Rodić, N., Zampella, J. G., Cornish, T. C., Wheelan, S. J., and Burns, K. H. (2013). Translocation junctions in *TCF3-PBX1* acute lymphoblastic leukemia/lymphoma cluster near transposable elements. *Mobile DNA*, *4*, 22.

Schmitz, R., Young, R. M., Ceribelli, M., Jhavar, S., Xiao, W., Zhang, M., Wright, G., Staudt, L. M. (2012). Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature*, *490*(7418), 116-120.

Sen, D., and Gilbert, W. (1988). Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature* 334(6180): 364-366

Sen, D., and Gilbert, W. (1990). A sodium-potassium switch in the formation of four-stranded G4-DNA. *Nature*, *344*(6265), 410-414.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308-311.

Spell, R. M., & Jinks-Robertson, S. (2004). Determination of mitotic recombination rates by fluctuation analysis in Saccharomyces cerevisiae. In *Genetic Recombination* (pp. 3-12). Humana Press.

Steininger, A., Möbs, M., Ullmann, R., Köchert, K., Kreher, S., Lamprecht, B., Anagnostopoulos, I., Assaf, C. (2011). Genomic loss of the putative tumor suppressor gene E2A in human lymphoma. *The Journal of experimental medicine*, *208*(8), 1585-1593.

Sun, D., and Hurley, L. H. (2010). Biochemical techniques for the characterization of G-quadruplex structures: EMSA, DMS footprinting, and DNA polymerase stop assay in *G-Quadruplex DNA* (pp. 65-79). *Humana Press*.

Tarsounas, M., Tijsterman, M. (2013). Genomes and G-quadruplexes: for better or for worse. *Journal of molecular biology* 425:4782-4789.

Tsai, A. G., Lu, H., Raghavan, S. C., Muschen, M., Hsieh, C. L., and Lieber, M. R. (2008). Human chromosomal translocations at CpG sites and a theoretical basis for their lineage and stage specificity. *Cell*, *135*(6), 1130-1142.

van Kregten, M., and Tijsterman, M. (2014). The repair of G-quadruplex-induced DNA damage. *Experimental cell research*, *329*(1), 178-183.

Vorlíčková, M., Kejnovská, I., Sagi, J., Renčiuk, D., Bednářová, K., Motlová, J., & Kypr, J. (2012). Circular dichroism and guanine quadruplexes. *Methods*, *57*(1), 64-75.

Wang, G., and Vasquez, K. M. (2004). Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(37), 13448-13453.

Weitzmann, M. N., Woodford, K. J., & Usdin, K. (1996). The development and use of a DNA polymerase arrest assay for the evaluation of parameters affecting intrastrand tetraplex formation. *Journal of Biological Chemistry*, *271*(34), 20958-20964.

Wiemels, J. L., Leonard, B. C., Wang, Y., Segal, M. R., Hunger, S. P., Smith, M. T., Crouse, V.,and Pine, S. R. (2002). Site-specific translocation and evidence of postnatal origin of the t (1; 19) E2A-*PBX1* fusion in childhood acute lymphoblastic leukemia. *Proceedings of the National Academy of Sciences*, *99*(23), 15101-15106.

Williamson, J. R., Raghuraman, M. K., and Cech, T. R. (1989). Monovalent cation-induced structure of telomeric DNA: the G-quartet model. *Cell*, *59*(5), 871-880.

Wu, Y., and Brosh, R. M. (2010). G-quadruplex nucleic acids and human disease. *FEBS journal*, *277*(17), 3470-3488.

Yadav, P., Harcy, V., Argueso, J. L., Dominska, M., Jinks-Robertson, S., and Kim, N. (2014). Topoisomerase I Plays a Critical Role in Suppressing Genome Instability at a Highly Transcribed G-Quadruplex-Forming Sequence. *PLoS genetics*, *10*(12), e1004839.

Zhao, J., Bacolla, A., Wang, G., and Vasquez, K. M. (2010). Non-B DNA structure-induced genetic instability and evolution. *Cellular and molecular life sciences*, *67*(1), 43-62.

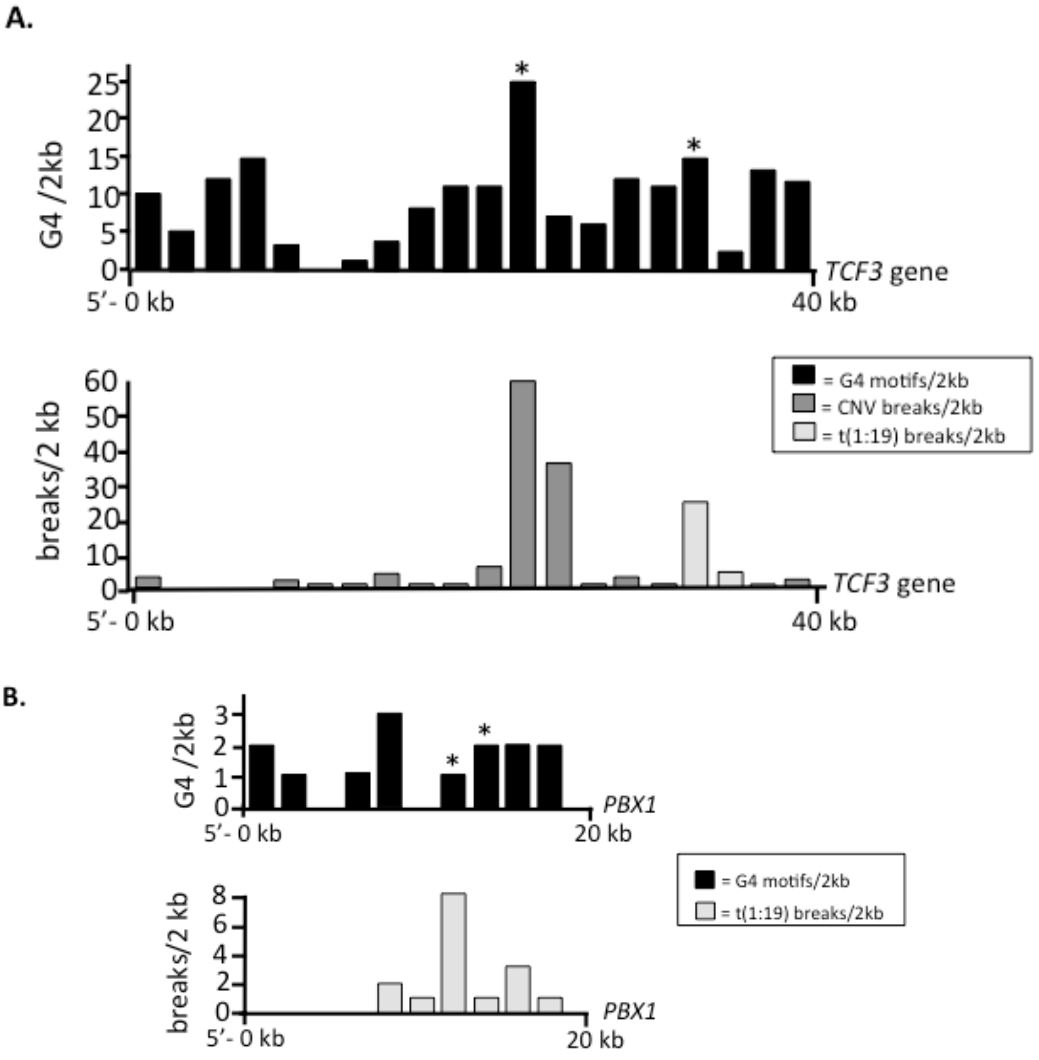

Figure 13. Genome Instability Coincides with G4 Motifs in *TCF3* and *PBX1*. (A and B) The number of individual non-overlapping G4 motifs (black) identified in 2 kb windows (Y-axis) graphed according to genetic position (X-axis) within the **(A)** *TCF3* gene or **(B)** 20 kb region of *PBX1* intron associated with the t(1;19) translocation, displayed 5' to 3' direction. Regions corresponding to motifs tested for G4 formation are indicated (*). The bar graphs below the *TCF3* and *PBX1* gene diagrams show the relative locations of break sites (light grey) and CNVs (dark grey) identified using 2 kb sequence windows.

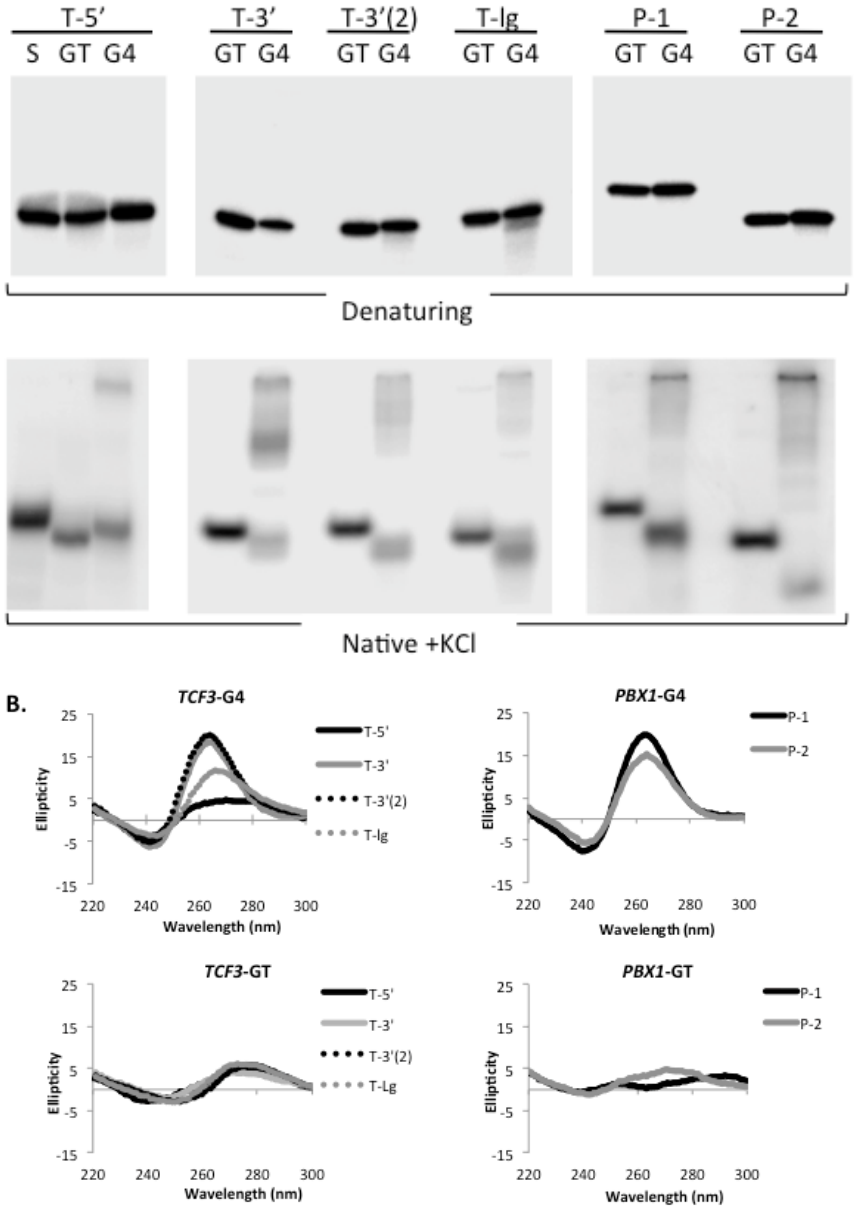

Figure 14. Sequences from *TCF3* and *PBX1* Adopt G4 Conformations in Solution. (A) Phosphorimages showing 5' end labeled guanine-rich oligonucleotides (G4) and corresponding thymine substituted controls (GT) resolved by PAGE. T-5' also includes an additional scrambled control (S), where nucleotides were reordered to remove the potential for stable hairpin folding. Migration of each radiolabeled oligonucleotide upon denaturing PAGE (top) and native PAGE (bottom) is shown. **(B)** CD spectra for *TCF3* (top left) and *PBX1* (top right) G4 motifs. CD spectra are shown for thymine-substituted (GT) controls for *TCF3* (bottom left) and *PBX1* motifs (bottom right).

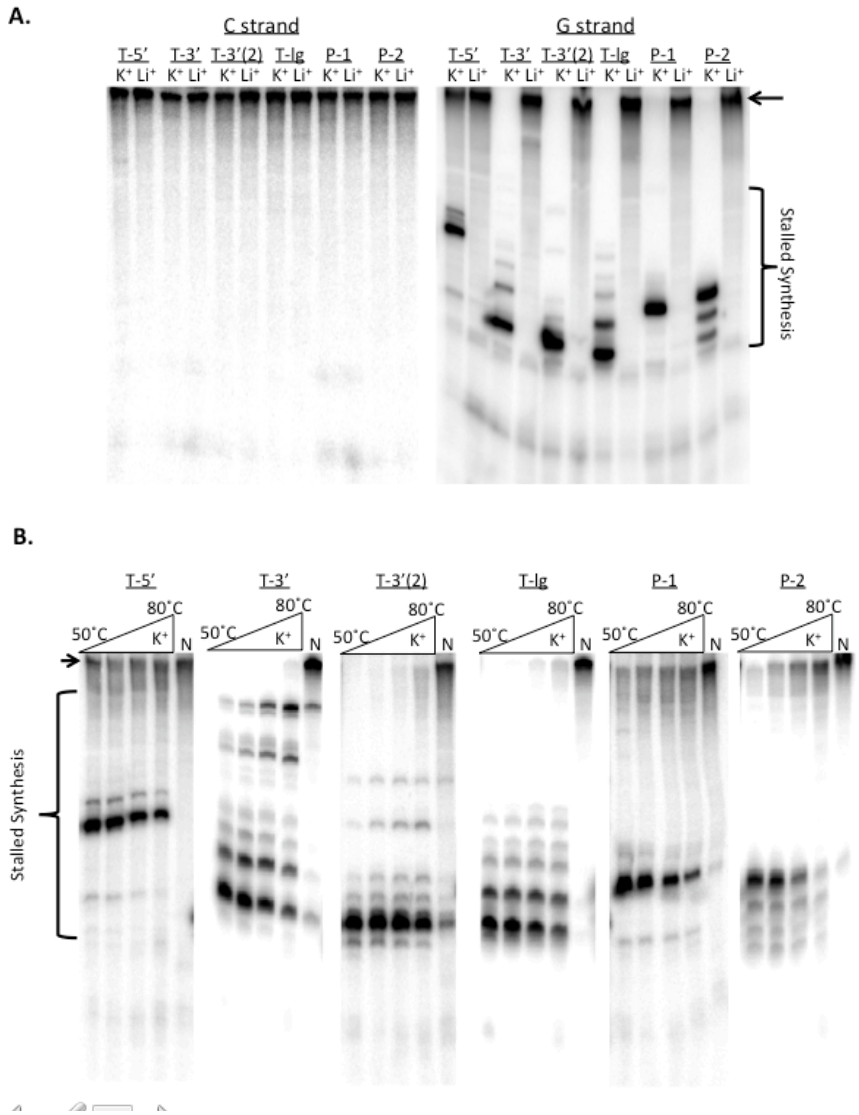

Figure 15. Guanine-Rich Templates From *TCF3* and *PBX1* Block DNA Synthesis *In Vitro*. (A) Klenow polymerase extension assays using templates from the cytosine-rich strand (C-strand, left) or guanine-rich strand (G-strand, right) for each G4 sequence motif, resolved by denaturing PAGE. Reactions were performed in G4-permissive salt conditions, KCL (K^+) or G4-disruptive salt conditions, LiCl (Li^+). Shown are bands for stalled DNA synthesis (bracket) or full-length extension (arrow). T-3' differs from the motif listed in Table 12, it includes 472 bp of surrounding genomic sequence (Figure 18). **(B)** Primer extension reactions used Taq polymerase across a temperature range (50°-80°C) in G4 permissive salt conditions, KCl (K^+), or G4 disruptive salt conditions, $(NH_4)_2SO_4$ (N) on guanine-rich templates from *TCF3* and *PBX1*. Bands corresponding to stalled DNA synthesis (bracket) or full-length extension (arrow) are shown.

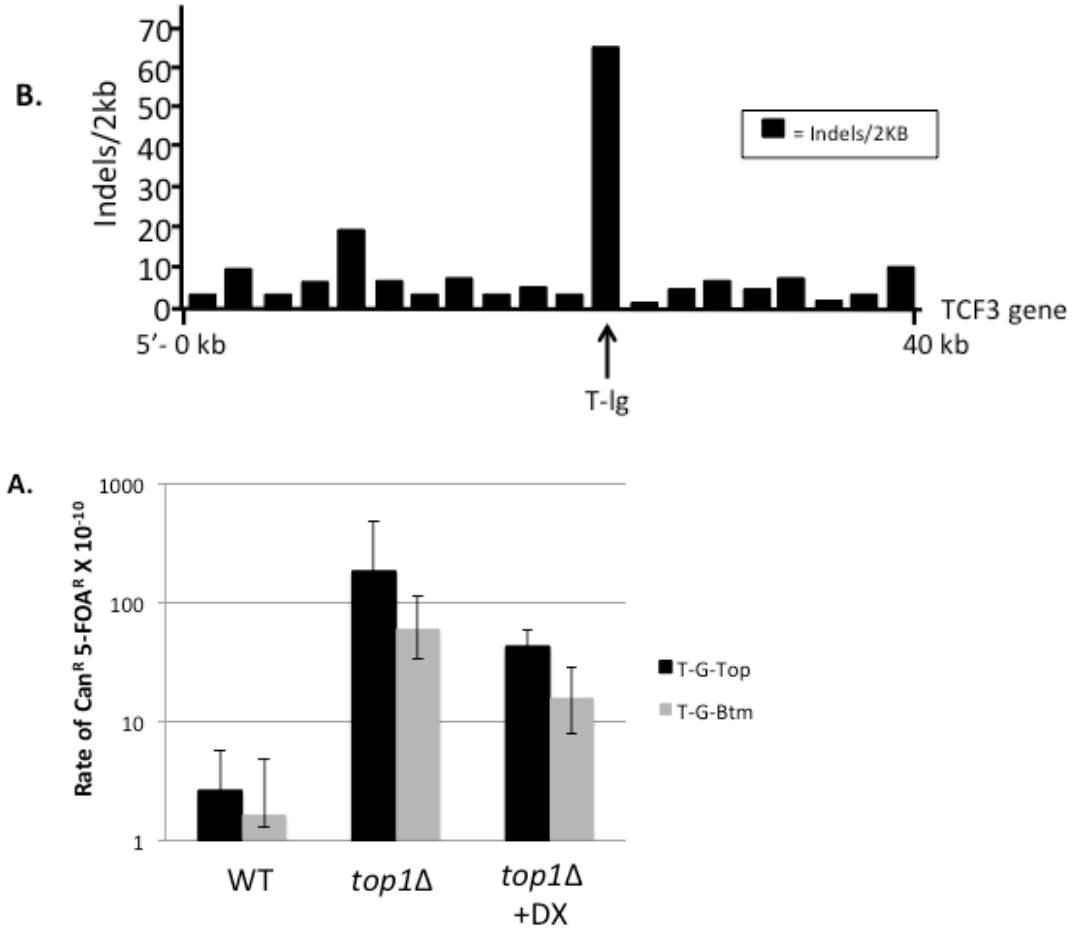

Figure 16. *TCF3* G4 Motifs Promote Genetic Instability *In Vivo*. (A) Rate of gross chromosomal rearrangements for the t(1;19) major break site sequence (T-3' G4 motif) in the *pTET-lys2 T-G-Top* (solid black bar) or *T-G-Btm* (grey bar) for WT or *top1* deletion yeast strains. Where indicated (+DX), transcription was repressed by addition of doxycycline to 2 mg/l. Error bars indicate 95% confidence intervals. **(B)** *TCF3* intron with high CNV (T-Ig) also shows high levels of small insertions and deletions (<100 bp). Insertion and deletions (indels) identified in *TCF3* using 2 kb sequence windows are mapped (indels / 2kb) according to genetic position.

Table 12. *TCF3* and *PBX1* G4 Sequences.
Names given for each sequence are in the far left column. Guanines that are predicted to be involved in G4 formation are bolded. QGRS scores are listed in the right column. The T-lg G4 sequence is not associated with t(1;19) break sites, rather it is a repeat motif located at the CNV site diagramed in Figure 1.

Name	Sequence	QGRS score
T-5'-G4	CCA**GGGG**ACACT**GGG**TGATGTCT**GGGG**ACATCTACAGTTGTCA**GGG**CTGA**GGGG**AGC	62
T-3'-G4	AGA**GGG**AGAGA**GGG**AA**GGGGGG**A**GGG**C**GGGG**CA**GGG**CAG	72
T-3'(2)-G4	A**GGG**AGT**GGGG**ACGTGAAT**GGGG**TGCGA**GGGG**C**GGGG**TG	101
T-Lg-G4	A**GGGGG**TGAGGC**GGG**AA**GGGG**ACAGCAGAACTCAC**GGGG**T	61
P-1-G4	GGT**GGGGG**CAGGTT**GGG**A**GGGG**AGGA**GGG**CAGATCTACA**GGG**A**GG**TGG	70
P-2-G4	GCT**GGGG**T**GGGG**A**GGG**AAGAGATGA**GGGGG**A**GGG**AGA	66

Table 13. All Oligonucleotide Sequences.
TCF3 and *PBX1* genomic sequences (G4) and companion controls (S and GT) used in G4 structure formation assays (center), QGRS scores (left). T-5'-scambled (S) is an additional control sequence that has a rearranged sequence to prevent hairpin or G4 structure formation.

Name	Sequence	QGRS score
T-5'-S	GACTGATCAGGTGAGCGGCGTAGTAGCTGGTGCTAAGGCGTGAGGAGGTGGCATCCC	0
T-5'-GT	CCAGTTGACACT**GGG**TGATGTCTGTTGACATCTACAGTTGTCAGTGCTGA**GGGG**AGC	0
T-5'-G4	CCA**GGGG**ACACT**GGG**TGATGTCT**GGGG**ACATCTACAGTTGTCA**GGG**CTGA**GGGG**AGC	62
T-3'-GT	AGAGTGAGAGAGTGAAGGTGTGAGTGCGTGGCAGTGCAG	0
T-3'-G4	AGA**GGG**AGAGA**GGG**AA**GGGGGG**A**GGG**C**GGGG**CA**GGG**CAG	72
T-lg-GT	A**GGTGG**TGAGGC**GTG**AA**GTGG**ACAGCAGAACTCAC**GTGG**T	0
T-lg-G4	A**GGGGG**TGAGGC**GGG**AA**GGGG**ACAGCAGAACTCAC**GGGG**T	61
T-3'(2)-GT	AGTGAGTGTTGACGTGAATGTTGTGCGAGTTTCGTTGTG	0
T-3'(2)-G4	A**GGG**AGT**GGGG**ACGTGAAT**GGGG**TGCGA**GGGG**C**GGGG**TG	101
P-1-GT	GGTGTTGGCAGGTTGTGAGTTGAGGAGTGCAGATCTACAGTGA**GGG**TGG	0
P-1-G4	GGT**GGGGG**CAGGTT**GGGA**GGGGAGGA**GGG**CAGATCTACA**GGGA**GGG**T**GG	70
P-2-GT	GCTGTTGTGTTGAGTGAAGAGATGAGTTGGA**GGG**AGA	0
P-2-G4	GCT**GGGG**T**GGGG**A**GGG**AAGAGATGA**GGGGG**A**GGG**AGA	66

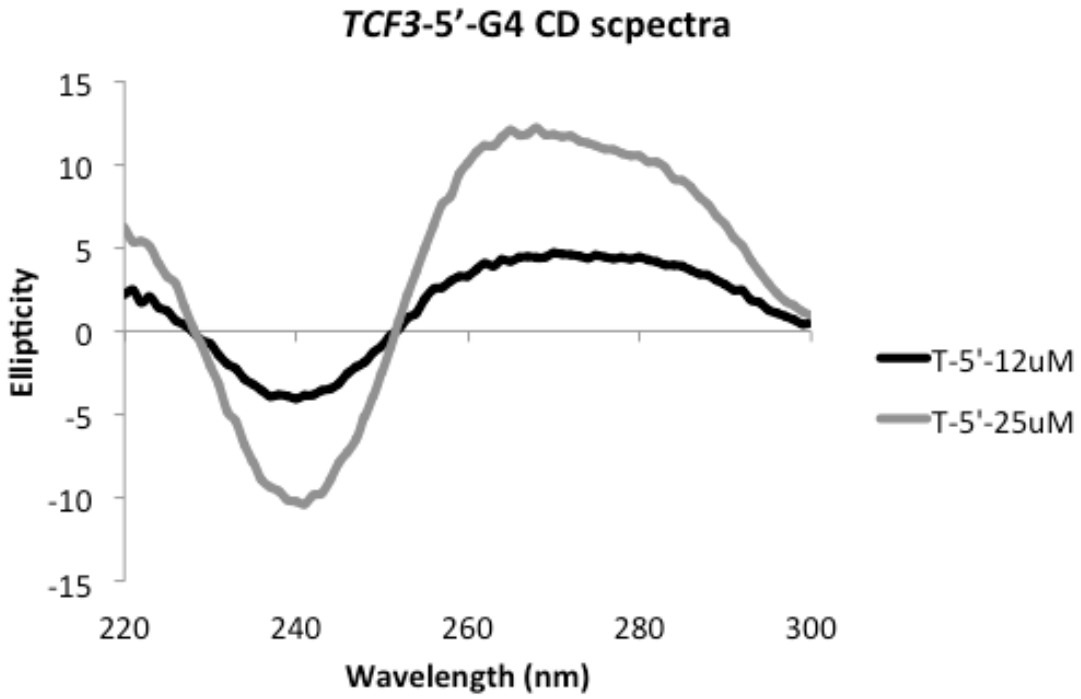

Figure 17. CD Spectra for T-5'-G4 CD spectra for T-5'-G4 are shown at 12 uM and 25 uM oligonucleotide concentrations.

T-5′-G4
GCCCTTCCA**GGGG**ACACT**GGG**TGATGTCT**GGGG**ACATCTACAGTTGTCA**GGG**CTGA**GGGG**AGCTCCTGGCATGGAGT**GG**
GTGGGGG**CCA**GGG**ATGCTGCTCAGCACCCTGCAGTGCCCAGGACGGCCCCACCCCAGAGAA**GGG**TCCGGACCCACAAA**G**
GGC

T-5′-GT
GCCCTTCCA**GGGG**ACACT**GGG**TGATGTCT**GGGG**ACATCTACAGTTGTCAGTGCTGAGGTGAGCTCCTGGCATGGAGTGT
GTGGTGGCCA**GGG**ATGCTGCTCAGCACCCTGCAGTGCCCAGGACGGCCCCACCCCAGAGAA**GGG**TCCGGACCCACAAA**G**
GGC

T-3′-G4
GCCCTT**GGGG**TTGGAGCAGAGTGAGGAGA**GGG**AGAGA**GGG**AA**GGGGGG**A**GGG**C**GGGG**CA**GGG**CAGGTCATGC**GGGG**CCT
TGT**GGG**CTGC**GGGG**AGGACTT**GGG**ATTTGGCCATGAGAAAGGTGGCAGCCGTGGA**GGG**CTGAGGA**GGG**AT**GGG**ACCTGA
CCCAGGTGCTCACAGATACCCTCTGGTGGCTGCTTGGAGGACAGACTATGAGTGGCACC**GGG**AGACCAAGGCAGAGGCC
ACA**GGG**CTGGTCCA**GGGG**CCTTGGA**GGGG**TGAGCAGT**GGG**T**GGG**CCCTGNATCTCCTGAAGGC**GGGG**CCCACAGGATTT
GTGATGGACAGGACTTGA**GGG**TGAGAGAAGGCA**GGG**TGGCTCT**GGG**ATTTCTGGCCCGAGCTGT**GGG**TGACACT**GGGG**C
T**GGG**AACTCCGTAAGGAA**GGGG**ACAAAGGAAAAGGTT**GGGG**ACAAATCTGGACCT**GGG**CCT**GGGG**ATTGTTAA**GGG**C

T-3′(2)-G4
GCCCTTCTGTC**GGGG**AA**GGG**T**GGGG**T**GGGG**C**GGGG**CAGGCACTCACCAGGCCGGAGACCCCCGTCGTAGCT**GGG**CGATA
AGGCACC**GGGGG**CTCCTGCTCGAGGCCACTGTGACGTTCCTGGAA**GGG**AGT**GGGG**ACGTGAAT**GGGG**TGCGA**GGGG**C**GG**
GGTGTAA**GGG**

T-lg-G4
T**GGGG**ATGAGGC**GGG**AA**GGGG**ACAGCAGAGCTCACA**GGGGG**TGAGGC**GGG**AA**GGGG**ACAGCAGAACTCAC**GGGG**TGAGG
C**GGG**AA**GGGG**ACAGCAGAGCTCACA**GGGGG**TGAGGCGGAAA**GGGG**

PBX1-1-G4
GCCCTTCTTCCACAGGT**GGGGG**CAGGTT**GGG**A**GGGG**AGGA**GGG**CAGATCTACA**GGG**A**GGG**TGGTCTTCAGATTTGGACA
ACAGTCCAA**GGG**CGCC

PBX1-1-GT
GCCCTTCTTCCACAGGTGTTGGCAGGTTGTGAGTTGAGGAGTGCAGATCTACAGTTAGGGTGGTCTTCAGATTTGGACA
ACAGTCCAAGGGCGCC

PBX1-2-G4
GCCCTTCCTATTTCAGAAAAAAGCTGCT**GGGG**T**GGGG**A**GGG**AAGAGATGA**GGGGG**A**GGG**AGAGAGCGCA**GGG**CACCCAT
CA**GGG**AAAA**GGG**C

PBX1-2-GT
GCCCTTCCTATTTCAGAAAAAAGCTGCTGGTGTGGTGAGTGAAGAGATGAGTTGGAGTGAGAGAGCGCAGTGCACCCAT
CAGTGAAAA**GGG**C

Figure 18. Sequences for Templates Used in Polymerase Extension Assays. Genome sequences PCR amplified or synthesized were cloned into pCR2.1, and include G4 motifs from *TCF3* or *PBX1* plus additional surrounding genomic sequence. Only the guanine-rich templates are shown, the complements are cytosine rich.

Figure 19. Cytosine-Rich Templates from *TCF3* and *PBX1* do not Stall Taq Polymerase. Taq polymerase extension assay using templates from the complementary cytosine-rich strand (C-strand) of each G4 sequence motif was resolved by denaturing PAGE and full-length extension products (arrow) are shown. Regardless of reaction conditions, G4-permissive salt conditions, KCL (K^+), or G4-disruptive salt conditions, $(NH_4)_2SO_4$ (N).

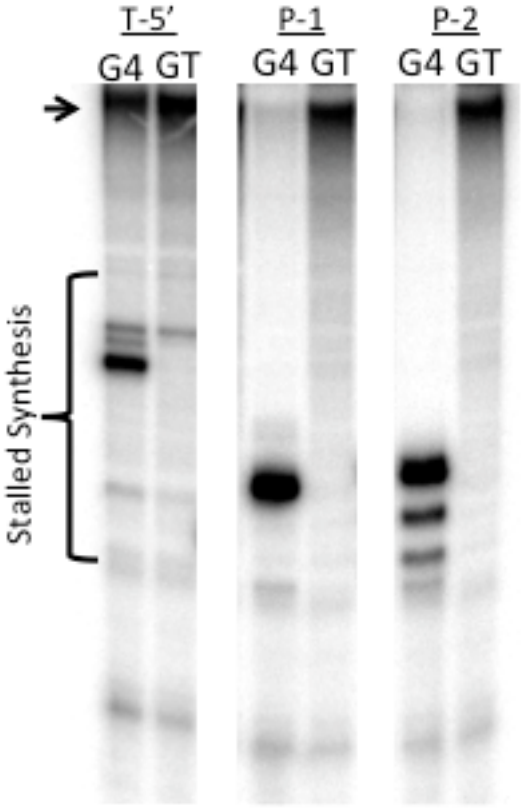

Figure 20. Polymerase Pausing *In Vitro* is Dependent on Guanine Triplets. Klenow polymerase extension assays using guanine-rich (G4), or guanine substituted (GT) templates for T-5', P-1 and P-2 sequences. Full primer extension products (arrow) and polymerase pausing, or stalled synthesis (bracket) are shown on the left.

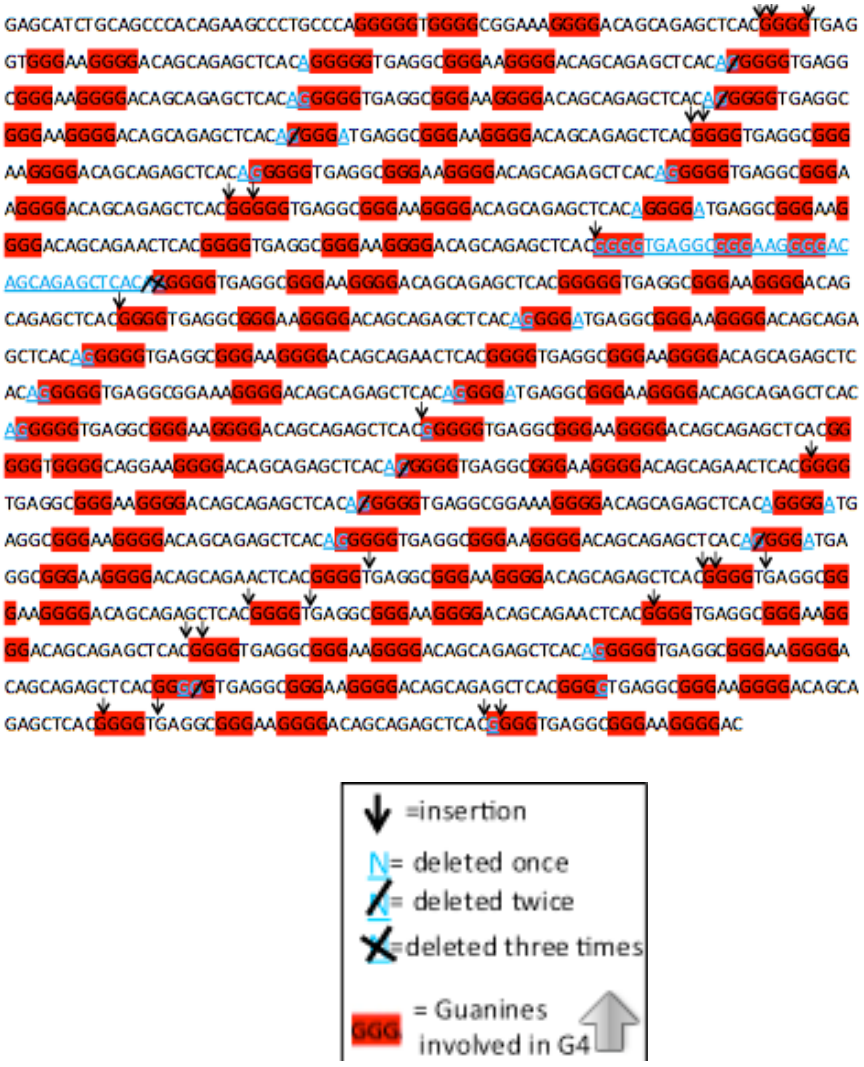

Figure 21. Sequence Location of T-lg Insertions and Deletions in Respect to Guanine Repeats. Ensembl77 was used to map location of insertions and deletions (indels) from the dbSNP database. Actual T-lg sequence shown with deletions and insertions identified denoted to the corresponding nucleotide.

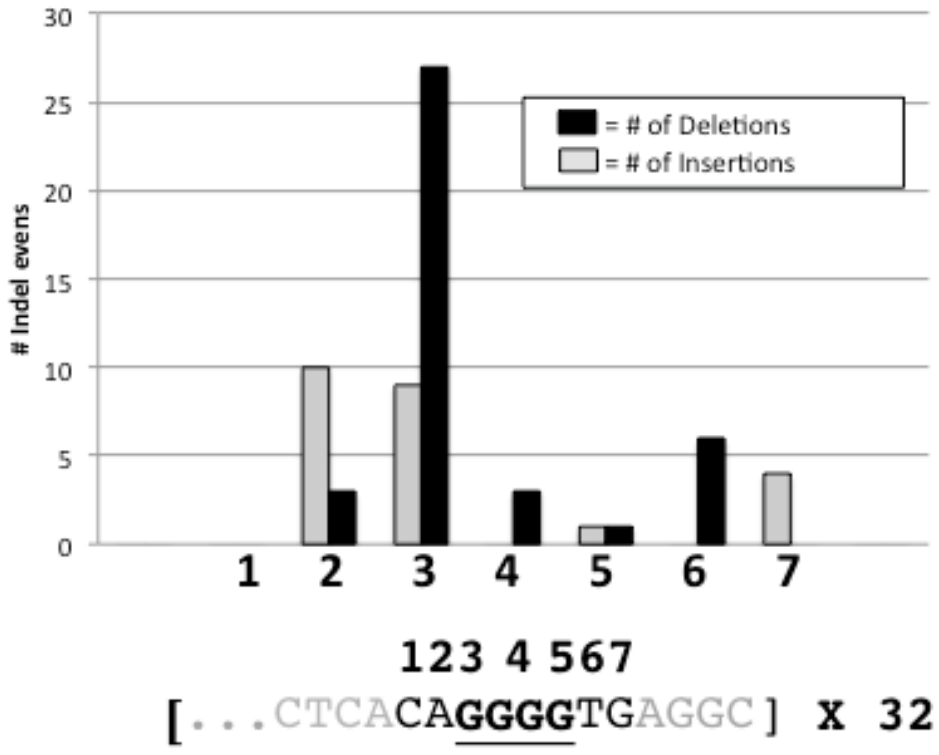

Figure 22. Graphed Location of T-lg Insertions and Deletions in Respect to Guanine Repeats.
Fine mapping of insertions and deletions with respect to the guanine repeat unit composing the T-lg intron. Numbers above the repeat motif correlate with numbers on the graph, depicting the location for each deletion (black bar) or insertion (grey bar) in the T-lg repeat motif.

CHAPTER IV

MISMATCH REPAIR AND G-QUADRUPLEX DNA;

A COMPLEX INTERACTION

Abstract

The mismatch repair system is required for proper genome maintenance and loss of pathway can lead to a mutator phenotype and cancer. Recently, the inability of DNA repair to operate at G4 structures has provided possible insights into the mechanism of site-specific instability at G-quadruplex (G4). However, the fidelity of the MMR system in G4 remains unknown. This is an important avenue of research considering the recognition protein of the MMR pathway (MutS) binds to G4 structures, most likely outside of MMR damage recognition. To investigate the MutS-G4 interaction I have employed a G4-phage infection system utilizing isogenic *E. coli* strains proficient and deficient in mismatch repair. Results provide evidence for a new role for mutS in G4 DNA metabolism. To inquire if this interaction had any effect on repair, I examined mutagenesis patterns inside and around the largest G4 sequences the human genome using sequence variation databases. I found a large increase in single nucleotide polymorphisms in LG4 and increased polymorphisms in adjacent microsatellites. The ability of the MMR system to directly repair a nucleotide mismatch was directly assayed using *in vitro* MMR reactions next to G4 sequences. Results demonstrate that MMR can be inhibited in an orientation and position dependent manner with respect to the G-rich sequence. Although, certain reaction conditions increased repair to levels previously unattained in G4 sequences. Together, these results suggest that G4 can lead to site-specific instability, possibly by inhibiting canonical MMR.

Introduction

The highly conserved mismatch repair (MMR) system is required for genome instability, and loss of pathway leads to a mutator phenotype and a predisposition to cancer (Jiricny, 2006; Li, 2008). MMR's primary function is to identify and correct replication errors (Li, 2008). Loss of the MMR pathway increases mutation rates up to 2400-fold (Tindall et al., 1998). Human MutSα, which is a heterodimer formed by pairing of the MSH2 and MSH6 proteins, is the primary recognition protein for of single base pair mismatches (Iaccarino et al., 1996; Fishel and Wilson, 1997). After damage recognition, MutSα hydrolyzes ATP to recruit downstream repair factors that remove the incorrect base. This is accomplished by the excision of the newly replicated strand by exonuclease activities (Li and Modrich, 1995). This part of the reaction involves other MMR proteins, such as MLH1 and PMS2. Re-synthesis through the gapped region by a polymerase, such as Pol delta, restores the parental strand information (Longley et al., 1997).

DNA repair pathways, including mismatch repair, may function outside of their known corrective roles in the presence of non-B form DNA structures, particularly G-quadruplex (G4 DNA). For instance, base excision repair removes modified bases from DNA, but in the presence of G4 structures classically defined activities are altered. Intriguingly, Neil1 glycosylase had increased activity for damage in G4 over duplex DNA (Zhou et al., 2013). In contrast, Ogg1 excision activity of 8-oxoG containing quadruplex is completely inhibited even though 8-oxoG is a high affinity substrate (Zhou et al., 2013). Direct reversal

repair of O^6alkyl guanine is also inhibited when those lesions are present in G4 DNA (Zhou et al., 2013).

Break repair activities at G4 sequences are also altered. Recently, results in *C. elegans* suggest that DNA breaks at G4 motifs were repaired by an error prone pathway that induced small insertions and deletions (Koole, et al., 2014). Repair was Pol theta-dependent and was outside homologous recombination and non-homologous end joining (Koole et al., 2014). Related to mismatch repair, we have previously found that *E. coli* MutS recognizes G4 DNA, but such binding does not initiate the classically defined excision repair portion of the pathway (Ehrat et al., 2012). At the Huntington locus, involved in Huntington's disease, MSH2/MSH3 complexes recognize imperfect hairpins, but ATPase activities are disrupted (Owen et al., 2005). If a mismatch was located within a quadruplex sequence, the formation of G4 DNA may inhibit portions of the pathway, or recruit binding by MutS proteins that could interfere with repair activities. The above examples illustrate an important point regarding G4 DNA in the genome; the increased mutagenesis characteristic of these sequences may be influenced in part by alternate or failed DNA correction activities at those loci.

This chapter will describe results regarding the activities of mismatch repair proteins in response to G4 DNA. We showed previously that bacterial and human MutS(α) both bind to G4 structures (Larson et al., 2005; Ehrat et al., 2012), and those results support a model whereby binding is outside of canonical MMR roles. MutS traditionally binds to a nucleotide mismatch with high affinity, and in the presence of ATP releases the mismatched substrate and signals

downstream repair (Li, 2008). When the amino acid motif in MutS required for ATP hydrolysis (F36) is substituted for alanine, or when ATP is added to the reaction, binding to G4 was not affected (Ehrat et al., 2012). Since ATP interactions are required for MutS activity in mismatch repair, these results indicate that the binding of G4 is not the same as mismatch binding. Considering the other parts of the pathway, when excision occurs through a guanine-rich sequence the transient loss of complementary base pairing could permit G4 formation. Those structures block DNA synthesis *in vitro* (see Chapter 2), and it is reasonable to predict that MMR could be blocked at that stage. In other words, I anticipate that G4 DNA is an impediment to normal mismatch repair, and this could be at more than one stage of the pathway.

I have taken a two-pronged approach to study the response of mismatch repair proteins to the presence of G4 DNA structures. This chapter will first describe experiments using a model G4 sequence, the Sγ3 immunoglobulin switch region, in a phage infection system utilizing isogenic *E. coli* strains proficient and deficient in mismatch repair. Results from this chapter describe a new role for MutS related to G4 DNA metabolism. In the second approach, I examine mutagenesis patterns in the human genome and assay mismatch repair functions in G4 sequences using a standard *in vitro* repair assay (Holmes et al., 1990; Thomas et al., 1991; Larson et al., 2002; Larson et al., 2008)**.** My results indicate that large G4 loci in the human genome are prone to small sequence variations, possibly from inhibition of MMR.

Materials and methods

Phage Assays

A fragment of human Ig Sγ3 sequence was cloned into pCR2.1 and then subcloned into M13mp18 (NEB, New England Biolabs, Ipswich, MA) as a *Xba*I and H*ind*III restriction fragment. This was done in two orientations resulting in phage containing either the C-rich strand (M13-C) or G-rich strand (M13-G). Phage cloning was verified by DNA sequencing. All plaque assays and phage purifications followed standard protocols. Larger volume phage stocks of wild-type and M13-G were created by infecting 500 µl of XL2 Blue (Stratagene, Allegient Technologies, Santa Clara, CA) at an OD of 0.5 with a single plaque, followed by culturing in 25 ml of LB overnight. Phages were concentrated by standard 2.5 M NaCl/20% PEG precipitation protocol (NEB), and then resuspended in 800 µl of TE buffer. Titers were determined by serial dilution and counting plaques. Appropriate volumes of either M13mp18 or M13-G were added to experimental NM522 to correspond to approximately 100 plaques for each plate. Plaque-forming efficiency for *mutS::Tn*10 NM522 (JW1) for M13mp18 or M13-G is presented relative to plaque-forming efficiency for NM522 (isogenic to JW1 and MutS proficient). JW1 cells were transformed with MutS F36A under control of the pTrc promoter in pTrcHIS2B or with pTrcHIS2B empty vector, and plaque-forming efficiency was relative to NM522 infection. In both empty vector and MutS F36A, expression from the pTrc promoter was induced by addition of 1 mM IPTG for 20 minutes prior to phage infection and plated on LB agar containing 1 mM IPTG.

Sequence Analysis

The output of Java LG4 identification program (chromosome and base pair location, Chapter 2) was used to map each individual LG4 location on Ensembl Release 69(hg19) (Flicek et al., 2011). Both adenine and thymine mononucleotide repeats over 12 bp within 2 kb from LG4s were identified using Microsoft Word search function, and subdivided into categories as either directly next to the LG4s (<1 kb) or further away (1-2 kb). The number of microsatellite deletion or insertion events was counted using the dbSNP database (Sherry et al., 2001) filter on Ensembl69 (Flicek et al., 2011) and normalized according to the sum of total mononucleotide base pairs.

MMR Substrate Prep

All clones used for MMR substrate prep were constructed by Topo cloning PCR amplified DNA into the pCR2.1 (Invitrogen) vector. For the assay of different mismatch locations, H*ind*III sites were removed and relocated to different parts of the vector and G4 sequence by site directed mutagenesis using heteroduplex primers and mismatch repair deficient E. coli (JW1) (Ehrat et al., 2012). All clones were verified sequencing before use in preparation of MMR substrates.

Preparations of MMR substrates was preformed as previously described (Larson et al.,2008). Briefly, closed circular single-stranded DNA was produced using M13K07 helper phage (NEB), according to the manufacturer's instructions. A heteroduplex 50-nucleotide primer was annealed to single-stranded templates to create a G-T mismatch at a *Hind*III restriction site. Next, Phusion polymerase

(NEB) extended this primer to form a complete, nicked circular heteroduplex molecule. Extension reaction were performed at 65°C for 45 minutes. Un-hybridized oligonucleotide, single-stranded DNA, and incomplete synthesis products were removed by BND cellulose (Sigma Aldrich), and repair substrates were purified by standard ethanol precipitation.

In Vitro MMR Reactions

In vitro MMR reactions used nuclear cell extracts from HeLa or Ramos cells as described previously (Larson et al.,2008, Holmes et al.,1990). Briefly, 100 ng of G-T heteroduplex repair substrate was incubated in 50 µg of nuclear extract, 20 mM Tris pH 7-7.9, 50 µg BSA, 50 100mM KCl, 5mM $MgCl_2$ 1mM glutathione, 1.5 mM ATP, and 0.1mM each dNTP for 25 minutes at 37°C. Reactions were terminated by addition of stop solution, 1:1 volume, containing 25mM EDTA, 1% SDS, 0.1 mg/ml Proteinase K, and incubated at 37°C for 15 minutes. Substrates and repair products were purified by phenol extraction followed by ethanol precipitation. G-T mismatch correction was then detected using *Hind*III/*Xmn*I, or *Hind*III/*Nco*I double restriction digestion, which contained RnaseA. Cleavage products of repair reactions and controls were resolved by 1% agarose gel electrophoresis containing ethidium bromide. Pixel values were quantified for each digestion band using ImageJ, and percentage repair calculated in excel using the following formula: sum of pixel value of bottom two bands divided by the sum of all three bands pixel value.

Statistical Analysis

All statistical analysis were preformed using StatPlus:macV5. Statistical analysis of SNPs used one-way ANOVAs. Comparisons of *in vitro* experimental repair levels to controls repair levels used an unpaired two-tailed T-test. *P*-values displayed on following graphs as follows (***=$p<.01$, *=$p<.05$).

Results

MutS is Required For Proper Infection of M13 Phage Encoding G4 DNA

Previous research in the Larson Lab revealed that purified bacterial MutS has high affinity for G4 DNA structures *in vitro*, and this is independent of DNA repair activation (Ehrat et al., 2012). This implies a model whereby MutS binding to G4 has a functional role outside of the MMR pathway. To test that hypothesis, I employed *E. coli* and a filamentous phage M13 infection assay. M13 is ideal for the study of DNA structures because the infectious particle contains a circular-single-stranded genome that is generated by rolling circular replication of closed circular duplex form. Single-strands are free to adopt non-B form conformations, which could interfere with phage replication. Previously, we showed that the Sγ3 G-rich sequence pauses Klenow polymerase in a K^+ dependent manner, as part of a standard primer extension assay (Chapter 2 Figure 9A) (Ehrat et al., 2012). That same sequence was used here. Sγ3 was cloned into M13 (M13-G) so that the phage strand will contain the guanine-rich sequence. M13-G infection rates of *E. coli* were compared to the parent molecule (M13mp18). We also compared infection rates for WT and MutS deficient strains.

I infected MutS proficient (NM522), deficient (JW1), or JW1 expressing MutS F36A from an inducible plasmid (JW1-MutS F36A). MutS F36A binds to G4, but not to G-T mismatches (Ehrat et al., 2012). I then asked if MutS, and thus mismatch repair, influences the efficiency of infection (M13-G). Phage infection success for the MutS defective strain (JW1) were measured by counting plaques and normalizing the numbers to an isogenic MutS proficient strain (NM522) (Figure 23A) (Ehrat et al., 2012).

Successful M13 phage infection of bacteria results in a plaque on LB agar plates. I first tittered phage stocks of M13-G and M13mp18 using NM522 and defined the volume required to generate ~100 plaques/plate for each stock. Using identical volumes and conditions, M13mp18 infection showed nearly equal plaques/plate for both NM522 and JW1 (Figure 23B), indicating that the MutS protein is not required for efficient infection by M13mp18 phage. In contrast, infection of JW1 with M13-G resulted in ~50% fewer plaques relative to NM522 infection, suggesting that disruption of MutS interferes with phage infection when those phage contain the Sγ3 sequence. This is most likely not associated with mismatch repair activities, at least in the classical sense, because expression of MutS F36A in JW1, a mutant defective in mismatch binding and repair but functional for G4 binding ability, resulted in near complete restoration of M13-G phage infection to that of NM522 (Figure 23A-B) (Ehrat et al., 2012). This supports a physiological role for MutS in responding to G4 DNA. However, it does not characterize a G4-specific pathway, and whatever activity of MutS is

responsible for the phenotype is not likely connected with strand excision and repair (Ehrat et al., 2012).

Microsatellite Instability and LG4

E. coli MutS and the human counterpart binds to G4 DNA *in vitro* (Ehrat et al., 2012; Larson et al.,2005) . This binding does not appear to signal mismatch repair, which is characterized by strand excision and resynthesis activities that result in the repair of mismatches. Therefore, it is reasonable to predict that G4 DNA may interfere with normal heteroduplex correction, perhaps explaining the higher levels of mutagenesis observed at some G-rich proto-oncogenes (Nambiar et al; 2011, 2013; Siddiqui-Jain et al., 2002; Sun et al., 2005; Arora et al., 2011; Cogoi and Xodo, 2006; Brooks et al., 2010). In hereditary forms of colon cancer (HNPCC), mismatch repair becomes inactivated and this is observed by microsatellite instability, or the increase of deletions and insertions in short tandem repeats (Liu et al., 1995; Peltomäki, 2001). Since mismatch repair corrects replication errors at microsatellites, loss of the pathway leads to sequence polymorphisms at those sites. In Chapter 2, I discuss a dataset of large G4 regions (>600 base pairs), which we predict form G4 DNA based on sequence composition (Chapter 2, subset shown Figure 24A). Within 2 kb of these sequences are hundreds of adenine and thymine mononucleotide repeats (>12 bp), providing a potential marker for mismatch repair activity. If mismatch repair defects are reflected by microsatellite polymorphisms (size changes), I

predict that the microsatellites closest to large G4 regions will contain more insertions and deletions than those further away.

I counted the number of mononucleotide repeats that contained a deletion or insertion sequence variation listed in the dbSNP database on Ensemble69 (Sherry et al., 2001). Normalized data of microsatellites 0-1 kb from the G4 regions were two-fold more likely to contain an insertion or deletion than microsatellites 1-2 kb away (Figure 24B). This was especially true for polymorphisms greater than 4 nucleotides (Figure 24B >4Δ). An increase in microsatellite instability as a function of distance from a large G4 repeat suggests that MMR function could be impaired at loci proximal to large G4 forming regions.

SNPs are Increased in LG4s

Single nucleotide polymorphisms (SNPs) are single nucleotide base pair differences of a specific nucleotide at a given location (eg. C→T). If MMR is disrupted, replication errors and SNPs increase (Modrich and Lahue, 1996). Although sequence-specific inhibition of MMR has not been documented, if that were to occur I would expect to see an increase in SNPs at that locus. However, this is an indirect measure and the only conclusion I can draw based on this analysis is on relative mutagenesis, and not reflective of any specific pathway.

I asked if LG4s contained an increase of SNPs compared to surrounding introns on the dbSNP database (Sherry et al., 2001). I found that SNPs were significantly increased in LG4s' first 200 bp both 5' and 3' by 50% (red bars) (p=0.00006) compared to surrounding introns (Figure 25A). However, there was

no significant increase in in the middle sections of LG4s (orange bar) (Figure 25A). The significant increase in only the first 200 bp 5' and 3' of LG4 prompted me to ask if shorter LG4s contain an increase in SNPs over longer LG4s. I found that the longest group of LG4s (3.5-4.5 kb) contained almost a two-fold increase (p=0.008) in SNP density compared to smaller regions (1.5-3.5 kb) (Figure 25B). This increase of SNP density in 3.5-4.5 kb LG4 suggest that these regions have higher levels of mutagenesis. However, this model cannot explain the sharp increase in SNPs in the first 200 bp of LG4s. It is plausible that G4 motifs flanked by "normal" sequences are prone to a higher rate of replication errors, or other DNA damage. Considering that multiple error prone translesion polymerases are needed for proper maintenance of G4 regions (Betous et al., 2009; Northam et al., 2014), their activity must also be considered. Indeed, it is likely that multiple DNA repair pathways, or their inhibition, at G4 DNA can result in mutagenesis. I next examined MMR activity *in vitro* using synthetic mismatched substrates and G4 DNA sequences.

In Vitro MMR Assays in G4 Sequences

Cell free assays have been developed for examining MMR *in vitro*, and were essential for elucidating the molecular mechanism (Holmes et al., 1990; Muster-Nassal and Kolodner, 1886; Su et al., 1988; Thomas et al., 1991; Zhang et al., 2005; Constantin et al., 2005). Human repair substrates require a site-specific mismatch, and nick to direct excision to one strand (Constantin et al., 2005). For our protocol, a nick is introduced as a consequence of producing the circular

mismatched molecules (Larson et al., 2008). First, an oligonucleotide is annealed to a closed circular single-stranded DNA template to produce a G-T heteroduplex at a H*ind*III restriction site. Complete extension by a high fidelity polymerase produces the MMR substrate (Larson et al., 2008) (Figure 26A). Incubation in human HeLa or Ramos cell nuclear extracts provide all of the required protein components for MMR, and reconstitution of the *Hind*III restriction site is then used as the diagnostic for correction (Figure 26B). We used site-directed mutagenesis to move the position of the *Hind*III (and mismatch) relative the G4 sequence.

To test the ability of G4 sequences to inhibit MMR, we created substrates where a *Hind*III restriction enzyme site was positioned 70 nucleotides 5', 10 nucleotides 5', and 10 nucleotides 3' of the Sγ3 sequence. We predict that a DNA excision initiating at the nick would extend into the guanine-rich sequence, which would allow for G4 formation (Figure 27A). After incubation with Ramos nuclear extract, repair of G-T mismatches positioned 70 nucleotides away from G4 was equal to control (non-G4) substrates (2.1) (results not shown). Conversely, repair is greatly reduced (60%) when the mismatch is 10 nucleotides 5', and only slightly reduced (16%) 10 nucleotides 3' when compared to controls (Figure 27B). The difference between repair of 5'10 and 3'10 nucleotide mismatches could be explained by the presence of G4 DNA, which could inhibit the re-synthesis stage of the repair reaction (G4-5'-10, Figure 27A).

Previous assays in our lab have shown that multiple sequences capable of G4 formation readily stall polymerase synthesis reactions *in vitro* (Williams et al.,

2015; Ehrat et al., 2012). Therefore it is possible that because of G4 formation in the excision tract, re-synthesis is inhibited during MMR. To test if re-synthesis through the excision tract could be responsible for 5' to 3' repair discrepancies, a *Hind*III site was cloned into the middle of Sγ3 in both orientations so that strand excision removes the complement to the C-rich strand (C$-Flank), or the complement to the G-rich strand (G4-Flank, Figure 28A). Repair of G4-flank would create a gapped intermediate capable of G4 formation. Repair efficiency was compared to substrates that do not have G4 forming sequences (substrate is called 2.1). In Ramos extracts there was a dramatic reduction in repair when re-synthesis of the excision tract used the Sγ3 G-rich strand as a template (G4-flank, Figure 28B). This is not simply due to the repeats because repair levels equaled that of the control when the proportionately repetitive complement (C$-Flank, Figure 28B) was used.

We did obtain conflicting results, complicating the interpretation of MMR experiments. Initially, repair reactions for the G4 flanking substrates were low (7%), but later assays on the same substrates resulted in increased repair (61%). This is higher than control reactions, and suggests activities outside of mismatch repair (G4-F Figure 28 vs. Figure 29). The increase in repair was also observed for G-T mismatches 10 nucleotides 5' of G4, but not to the degree as G-T mismatches directly in the middle of G4. While the reasons for the different results are not clear, I did change reaction buffer conditions that could have influenced activities. Intriguingly, repair of 5'70, 2.1, and all C-rich repair

substrates were not affected by the reaction condition change (Figure 29 and results not shown).

The conflicting results that I obtained were addressed by changing experimental conditions. HeLa cells were used to produce the nuclear extract, the G4 motif was changed, and fresh ATP and dNTPs were made. None of these modifications altered repair levels. The pH of the reaction may influence the repair outcomes. Surprisingly, using HeLa extract, I was able to increase repair of the TCF3 G4-flank substrate at higher pH (7.6) (results not shown). Considering that G4 can forms at neutral pH (Yan et al., 2013), this small change in pH most likely had no effect on G4 formation. It seems more likely that a protein activity that is sensitive to pH has been altered, and this resulted in efficient excision and resynthesis (*Hind*III site reconstitution). Interestingly, these observations suggest that the activities I am observing enhance the resolution of G4 DNA, or otherwise allow for synthesis activities through the mismatch site. Further research is needed to clarify the way mismatch repair operates in the context of G4 sequences.

Discussion

Previous research has demonstrated that MutS specifically binds to G4 DNA *in vitro*, and this likely occurs outside of mismatch correction activities (Ehrat et al., 2012; Larson et al., 2005). Mismatch repair proteins function during replication (Ross-Macdonald and Roeder, 1994; Junop et al., 2003), so it is likely that MutS-related activities are focused at G4 DNA folded during replication or

recombination. However, MutS could also interact with G4 structures during the resynthesis stages of mismatch repair. Regardless, phage assays suggest that MutS facilitates the resolution of secondary structures (Figure 23) and informatic analysis suggest MMR is reduced inside and near G4 sequences (Figure 24).

Since the MutS mutant strain (JW1) and NM522 were infected equally well with M13mp18, but not M13-G, MutS may be required for proper phage maturation when its genome contains G4 sequences. Since the infectious particle of M13 contains a circular single stranded genome (Marvin, 1998), G4 DNA likely presents a physical block to phage replication. Stalled replication could lead to stalled replication forks, DNA breaks, and an SOS response. MutS binding may release that stress, potentially by enhancing G4 unwinding. This possibility has some support in the literature. In humans, MSH2/MSH6 complexes bind to G4 (Larson et al., 2005), and may help recruit helicase activities such as BLM (Pedrazzi et al., 2003). Either way, the results I present here show that MutS in *E. coli* has an influence on phage infection success when those viruses contain G-rich sequences that are supportive of G4 structures.

This chapter presents evidence that there is a biological response of bacterial MutS to sequences containing G4 motifs. This is the first cellular evidence suggesting that MutS proteins have some functional roles in responding to G4 DNA. MutS homologs have already been shown to have functions outside of MMR, so this may not be surprising. For instance, MSH2/MSH6 complexes may bind to cisplatin adducts and directly signal cell death (Wu et al., 1999; Bellacosa, 2001). In addition, MSH2/MSH3 proteins bind to hairpin structures

formed in Huntington's disease loci, and that may stabilize the structures as part of the causative event leading to repeat expansions (Lang et al., 2011; Owen et al., 2005). Interestingly, that binding is also un-responsive to ATP (Tome et al., 2009), suggesting that this is a widely-shared characteristic of the MutS homologs; structure binding is not the same as mismatch binding, and it elicits a yet to be defined cellular response. Whatever this response is, it cannot be MMR because ATP hydrolysis by MutS proteins is a required pre-requisite. Future research is needed in deciphering the precise role of MutS homologs in DNA structure interactions, and roles for other factors of the MMR pathway. One possibility is the recruitment of structure-resolution activities.

There is some evidence that MutS homologs recruit activities involved in resolving G4 DNA. RecQ is a G4 helicase whose activities are highly conserved with five homologs in humans and include BLM, FANCJ, RECQL1, RECQL4, and RECQL5 (Hickson, 2003). Interestingly, MutS has been shown to interact with 3' to 5' helicase BLM (Pedrazzi et al., 2003) and inhibit the activity of 5'-3' helicase FANCJ on unwinding G4 DNA (Wu et al., 2008). If MutS homologs bind to G4 DNA in order to facilitate its resolution during replication, one could predict helicase activities that favor the progression of replication (i.e., 3'-5'). DNA is synthesized by polymerases reading the template strand 3' to 5', and adding nucleotides on the 3' end of the newly synthesized strand. If replication stalled at a G4 structure on the template strand, the most efficient way of structure resolution would be in a 3'-5' direction, or in the direction with DNA synthesis. Alternatively, resolution starting from the 5' end of G4 would create an open and

unprotected single-stranded region, which would be more prone to damage. The notion of MutS binding to G4 structures and promoting replication progression through helicase recruitment fits previously described interactions of MutS homologs blocking 5'-3' FANCJ activity yet still allowing for 3'-5' BLM helicase activity (Model, Figure 30).

Even if MutS homologs bind to G4 and facilitates unwinding activities, that function could come at a cost to replication fidelity. MutS proteins help reduce replication errors by first recognizing mismatches and then initiating events leading to excision activities directed to the daughter strand (Li, 2008). If bound to G4 DNA instead, MutS homolog activity would be titrated away from any nearby mismatches. Alternatively, G4 could actively recruit repair factors to the area, and in turn increase repair levels in adjacent regions. That possibility seems less likely however because MutS and MSH2/MSH6 both fail to release from G4 DNA in vitro when ATP is included in the reaction (Ehrat et al., 2012; Larson et al., 2005). My bioinformatic analysis of microsatellites next to large G4 sequences does suggest that MMR frequently fails near G4 sequences (Figure 24). Although, the severity of mismatch repair inhibition in regions adjacent to G4 is most likely not as severe as inhibition of repair inside G4 repeats. This notion is supported by an increase of SNPs inside large G4 regions and not in adjacent sequences (Figure 25). In further support, mismatches 10 bp away from G4 were only slightly inhibited *in vitro* compared to mismatches flanked by G4 (Figure 27-28). Therefore, it is likely that mismatch repair is only slightly reduced in regions adjacent to G4 which manifest as increased variation in microsatellites

only; motifs especially sensitive to loss of MMR. The increase in small insertions and deletions in large G4 introns (Figure 25, Chapter 2) likely reflects a general trend toward instability that involves multiple repair activities. Although the mechanisms are unclear, it is plausible that SNPs, insertions, and deletions are induced from translesion polymerase activity, an increase in double strand break repair, or inability to properly repair nucleotide damage. The resulting mutagenesis from such otherwise unrelated repair pathways fits with the bioinformatics data.

Certainly *in vitro* MMR reactions support the notion that repair does not always function properly in G4 regions. During repair of substrates containing the guanine-rich DNA on the non-nicked strand, the excision tract creates a gapped substrate that can support G4 formation. Although results were conflicting, every assay displayed differential repair when conditions allowed G4 formation during repair intermediates. Because G4 can only form after strand excision, structure formation would most likely inhibit gap re-synthesis. Therefore it is possible that alternative DNA polymerases are necessary for strand re-synthesis through G4. This could be assayed by addition of aphidicolin, a traditional repair polymerase inhibitor, to MMR reactions (Crute et al., 1986). The use of different MMR deficient cell extracts could also be used to determine if alternative repair proteins are involved in G4 repair. Movement to a simpler model organism and development of an *in vitro* repair assay could aid in dissecting the molecular biology of MMR in G4 DNA. Although a difficult task, deciphering the complex

interaction of the MMR system and G4 could provide insights into site-specific genetic instability at G-rich repeats.

References

Arora, A., Suess, B. (2011). An RNA G-quadruplex in the 3'UTR of the proto-oncogene PIM1 represses translation. RNA biology, 8:802-805.

Bellacosa, A. (2001). Functional interactions and signaling properties of mammalian DNA mismatch repair proteins. *Cell death and differentiati*on, 8(11), 1076-1092.

Bétous, R., Rey, L., Wang, G., Pillaire, M. J., Puget, N., Selves, J., & Hoffmann, J. S. (2009). Role of TLS DNA polymerases eta and kappa in processing naturally occurring structured DNA in human cells. *Molecular carcinogenesis*, *48*(4), 369-378.

Brooks, T.A., Kendrick, S., Hurley, L. (2010). Making sense of G-quadruplex and i- motif functions in oncogene promoters*. Febs Journal* 277:3459-3469.

Cogoi, S., & Xodo, L. E. (2006). G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic Acids Res*earch 34:2536-2549.

Constantin, N., Dzantiev, L., Kadyrov, F. A., & Modrich, P. (2005). Human mismatch repair reconstitution of a nick-directed bidirectional reaction. *Journal of Biological Chemistry 280*:39752-39761.

Crute, J. J., Wahl, A. F., & Bambara, R. A. (1986). Purification and characterization of two new high molecular weight forms of DNA polymerase. delta. *Biochemistry*, 25(1), 26-36.

Ehrat, E. A., Johnson, B. R., Williams, J. D., Borchert, G. M., & Larson, E. D. (2012). G-quadruplex recognition activities of E. Coli MutS. *BMC molecular biology*, 13(1), 23.

Fishel, R., & Wilson, T. (1997). MutS homologs in mammalian cells. *Current opinion in genetics & development*, 7(1), 105-113.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Hubbard, T. J. (2011). Ensembl 2012. *Nucleic Acids Res*earch, gkr991.

Hickson, I.D., (2003). RecQ helicases: caretakers of the genome. *Nature Reviews Cancer 3*:169-178.

Holmes, J., Clark, S., Modrich, P. (1990). Strand-specific mismatch correction in nuclear extracts of human and DrosopHila melanogaster cell lines. *PNAS 87*:5837-5841.

Iaccarino, I., Palombo, F., Drummond, J., Totty, N. F., Hsuan, J. J., Modrich, P., & Jiricny, J. (1996). MSH6, a Saccharomyces cerevisiae protein that binds to mismatches as a heterodimer with MSH2. Current Biology, 6(4), 484-486.

Jiricny, J. (2006). The multifaceted mismatch-repair system. *Nature Reviews Molecular Cell Biology 7*:335-346.

Junop, M. S., Yang, W., Funchain, P., Clendenin, W., & Miller, J. H. (2003). In vitro and in vivo studies of MutS, MutL and MutH mutants: correlation of mismatch repair and DNA recombination. *DNA repair 2*:387-405.

Koole, W., van Schendel, R., Karambelas, A. E., van Heteren, J. T., Okihara, K. L., & Tijsterman, M. (2014). A Polymerase Theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites. *Nature communications* 5.

Lang, W. H., Coats, J. E., Majka, J., Hura, G. L., Lin, Y., Rasnik, I., & McMurray, C. T. (2011). Conformational trapping of mismatch recognition complex MSH2/MSH3 on repair-resistant DNA loops. *Proceedings of the National Academy of Sciences*, 108(42), E837-E844.

Larson, E. D., Nickens, D., & Drummond, J. T. (2002). Construction and characterization of mismatch-containing circular DNA molecules competent for assessment of nick-directed human mismatch repair in vitro. *Nucleic Acids Research 30*:e14-e14.

Larson, E.D., Iams, K., Drummond, J.T. (2003). Strand-specific processing of 8-oxoguanine by the human mismatch repair pathway: inefficient removal of 8-oxoguanine paired with adenine or cytosine. *DNA repair 2*:1199-1210.

Larson, E.D., Duquette, M.L., Cummings, W.J., Streiff, R.J., Maizels, N. (2005). MutSα binds to and promotes synapsis of transcriptionally activated immunoglobulin switch regions. *Current biology 15*: 470-474.

Larson, E.D, Bednarski, D.W., Maizels, N. (2008). High-fidelity correction of genomic uracil by human mismatch repair activities. *BMC molecular biology 9*:94.

Li, G.M. (2008). Mechanisms and functions of DNA mismatch repair. *Cell research 18*: 85-98.

Li, G. M., & Modrich, P. (1995). Restoration of mismatch repair to nuclear extracts of H6 colorectal tumor cells by a heterodimer of human MutL homologs. *Proceedings of the National Academy of Sciences*, 92(6), 1950-1954.

Liu, B., Nicolaides, N. C., Markowitz, S., Willson, J. K., Parsons, R. E., Jen, J., . Vogelstein, B. (1995). Mismatch repair gene defects in sporadic colorectal cancers with microsatellite instability. *Nature genetics*, *9*(1), 48-55.

Longley, M.J., Pierce, A.J., Modrich, P. (1997). DNA polymerase δ is required for human mismatch repair in vitro. *Journal of Biological Chemistry 272:*10917-10921.

Maizels, N. (2012). G4 motifs in human genes. *Annals of the New York Academy of Sciences* 1267: 53-60.

Marvin, D. A. (1998). Filamentous phage structure, infection and assembly. *Current opinion in structural biology*, 8(2), 150-158.

McMurray, C.T. (2010). Mechanisms of trinucleotide repeat instability during human development. *Nature Reviews Genetics* 11:786-799.

Modrich, P., & Lahue, R. (1996). Mismatch repair in replication fidelity, genetic recombination, and cancer biology. *Annual review of biochemistry*, *65*(1), 101-133.

Muster-Nassal, C., Kolodner, R. (1986). Mismatch correction catalyzed by cell-free extracts of Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences 83*:7618-7622.

Nambiar, M., Goldsmith, G., Moorthy, B. T., Lieber, M. R., Joshi, M. V., Choudhary, B., Raghavan, S. C. (2011). Formation of a G-quadruplex at the BCL2 major breakpoint region of the t (14; 18) translocation in follicular lymphoma. *Nucleic Acids Res*earch 39:936-948.

Nambiar, M., Srivastava, M., Gopalakrishnan, V., Sankaran, S. K., & Raghavan, S. C. (2013). G-quadruplex structures formed at the HOX11 breakpoint region contribute to its fragility during t (10; 14) translocation in T-cell leukemia. *Molecular and cellular biology*, 33(21), 4266-4281.

Northam, M. R., Moore, E. A., Mertz, T. M., Binz, S. K., Stith, C. M., Stepchenkova, E. I., Shcherbakova, P. V. (2014). DNA polymerases ζ and Rev1 mediate error-prone bypass of non-B DNA structures. *Nucleic Acids Res*earch, 42(1), 290-306.

Owen, B. A., Yang, Z., Lai, M., Gajek, M., Badger, J. D., Hayes, J. J., McMurray, C. T. (2005). (CAG) n-hairpin DNA binds to Msh2–Msh3 and changes properties of mismatch recognition. *Nature structural & molecular biology*, *12*(8), 663-670.

Pedrazzi, G., Bachrati, C. Z., Selak, N., Studer, I., Petkovic, M., Hickson, I. D., ... & Stagljar, I. (2003). The Bloom's syndrome helicase interacts directly with the human DNA mismatch repair protein hMSH6. *Biological chemistry 384*:1155-1164.

Peltomäki, P. (2001). Deficient DNA mismatch repair: a common etiologic factor for colon cancer. *Human Molecular Genetics*, *10*:735-740.

Ross-Macdonald, P., Roeder, G.S. (1994). Mutation of a meiosis-specific MutS homolog decreases crossing over but not mismatch correction. *Cell 79*:1069-1080.

Siddiqui-Jain, A., Grand, C. L., Bearss, D. J., & Hurley, L. H. (2002). Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *PNAS* 99:11593-11598.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research 29*:308-311.

Su, S.S., Lahue, R.S., Au, K.G., Modrich, P. (1988). Mispair specificity of methyl-directed DNA mismatch correction in vitro. *Journal of Biological Chemistry 263*:6829-6835.

Sun, D., Guo, K., Rusche, J.J., Hurley, L.H. (2005). Facilitation of a structural transition in the polypurine/polypyrimidine tract within the proximal promoter region of the human VEGF gene by the presence of potassium and G-quadruplex-interactive agents. *Nucleic Acids Res*earch 33:6070-6080.

Tarsounas, M., Tijsterman, M. (2013).Genomes and G-quadruplexes: for better or for worse. *Journal of molecular biology* 425:4782-4789.

Thomas, D.C., Roberts, J.D., Kunkel, T.A. (1991). Heteroduplex repair in extracts of human HeLa cells. *Journal of Biological Chemistry 266*:3744-3751.

Tindall, K.R., Glaab, W.E., Umar, A., Risinger, J.I., Koi, M., Barrett, J.C., Kunkel, T.A. (1998). Complementation of mismatch repair gene defects by chromosome transfer. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 402*:15-22.

Tomé, S., Holt, I., Edelmann, W., Morris, G. E., Munnich, A., Pearson, C. E., & Gourdon, G. (2009). MSH2 ATPase domain mutation affects CTG• CAG repeat instability in transgenic mice. *PLoS genetics*, 5(5), e1000482.

van Kregten, M., Tijsterman, M. (2014). The repair of G-quadruplex-induced DNA damage. *Experimental cell research* 329:178-183.

Williams, J.D., Fleetwood, S., Berroyer, A., Kim, N., Larson, E.D. (2015). F Formation of G-quadruplex DNA influences the genetic stability of human TCF3 (E2A). *Frontiers Journal* Submitted February

Wu, J., Gu, L., Wang, H., Geacintov, N. E., & Li, G. M. (1999). Mismatch repair processing of carcinogen-DNA adducts triggers apoptosis. *Molecular and cellular biology*, 19(12), 8292-8301.

Wu, Y., Shin-ya, K., Brosh, R.M. (2008). FANCJ helicase defective in Fanconia anemia and breast cancer unwinds G-quadruplex DNA to defend genomic stability. *Molecular and cellular biology 28:*4116-4128.

Yan, Y.Y., Tan, J.H., Lu, Y.J., Yan, S.C., Wong, K.Y., Li, D., Huang, Z.S. (2013). G-Quadruplex conformational change driven by pH variation with potential application as a nanoswitch. *Biochimica et BiopHysica Acta (BBA)-General Subjects 1830*:4935-4942.

Zhang, Y., Yuan, F., Presnell, S.R., Tian, K., Gao, Y., Tomkinson, A.E., Li, G.M. (2005). Reconstitution of 5′-directed human mismatch repair in a purified system *Cell* 122: 693-705.

Zhou, J., Liu, M., Fleming, A. M., Burrows, C. J., & Wallace, S. S. (2013). Neil3 and NEIL1 DNA glycosylases remove oxidative damages from quadruplex DNA and exhibit preferences for lesions in the telomeric sequence context. *Journal of Biological Chemistry*, 288(38), 27263-27272.

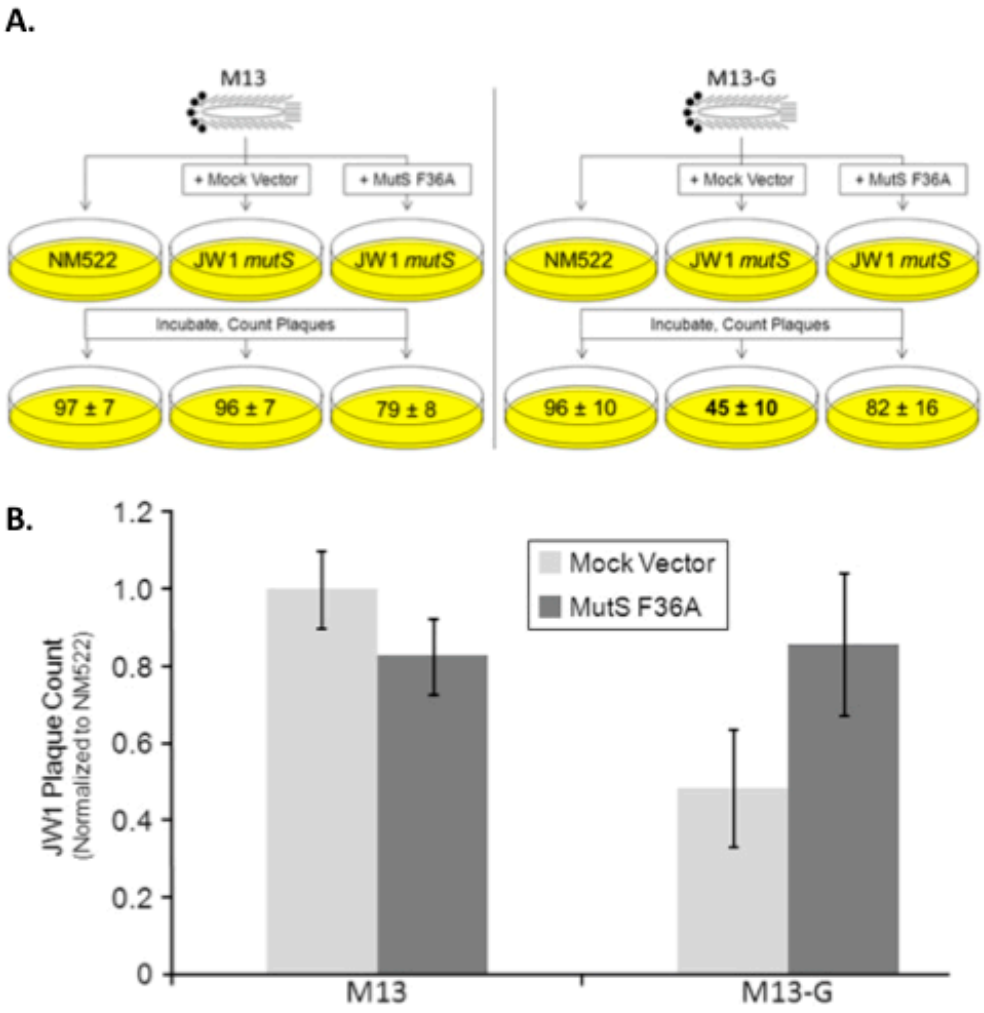

Figure 23. MutS F36A Facilitates Efficient Infection by G-Rich M13 Phage. To ask if MutS G4 binding activity influences phage infection success, I examined the abilities of a M13 variant, M13-G (M13 with Sγ3) and its parent molecule (M13mp18) to infect bacteria in the presence or absence of MutS expression. **(A)** Cartoon depicting plaque assay methodology. MutS proficient (NM522), deficient (NM522 mutS::TN10 (JW1)) transformed with empty (Mock Vector), or JW1 expressing MutS F36A from "MutS F36A" vector were infected to ask if MutS influences infection efficiency when M13 harbors a G4 competent sequence. Phage infection success for MutS defective strain was measured by counting plaques/plate. **(B)** Graph depicting results of assay diagrammed in (A). Phage infection success for the JW1 MutS defective strain was measured by counting plaques then normalizing to the isogenic MutS proficient strain NM522 (n = 6). M13-G infection rates of F36A (dark gray) versus mock vector (light gray) suggest MutS was required for proper M13-G infection outside of its role in MMR.

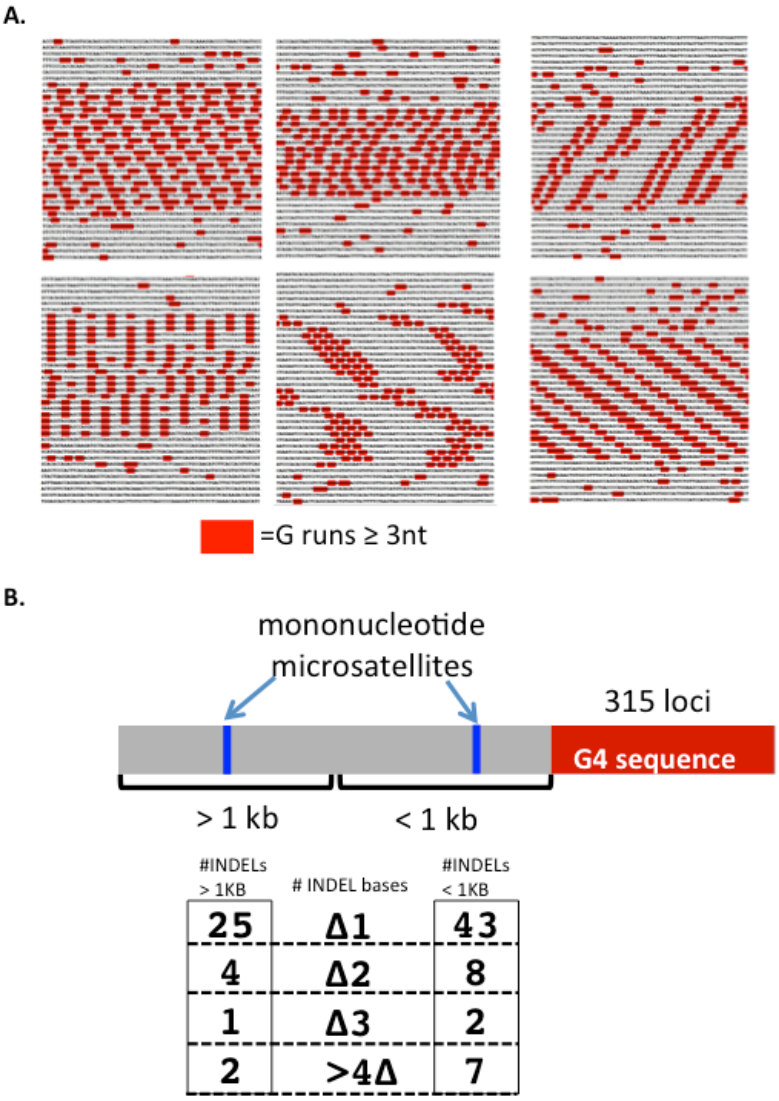

Figure 24. Mononucleotide Repeats are More Prone to Deletions and Insertions when Directly Next to LG4s. To investigate if MMR fidelity is potentially inhibited by MutS' high affinity for G4, insertions and deletions (indels) from the dbSNP database were harvested for mononucleotide repeats directly surrounding (<2kb) LG4s on Ensembl69. **(A)** Subset of LG4 sequences used in computational analysis of A and T microsatellites are shown. Distance of microsatellite from LG4 was calculated from the end of the G-triplet (red) rich repetitive sequence. **(B)** Schematic diagram (top) and corresponding normalized data (bottom) of the two groups of microsatellites analyzed, 0-1kb or >1kb. However, 5' or 3' microsatellite locations with respect to the LG4 were not differentiated in this analysis. Below, the normalized numbers of microsatellites containing indels of various sizes (middle column) are listed.

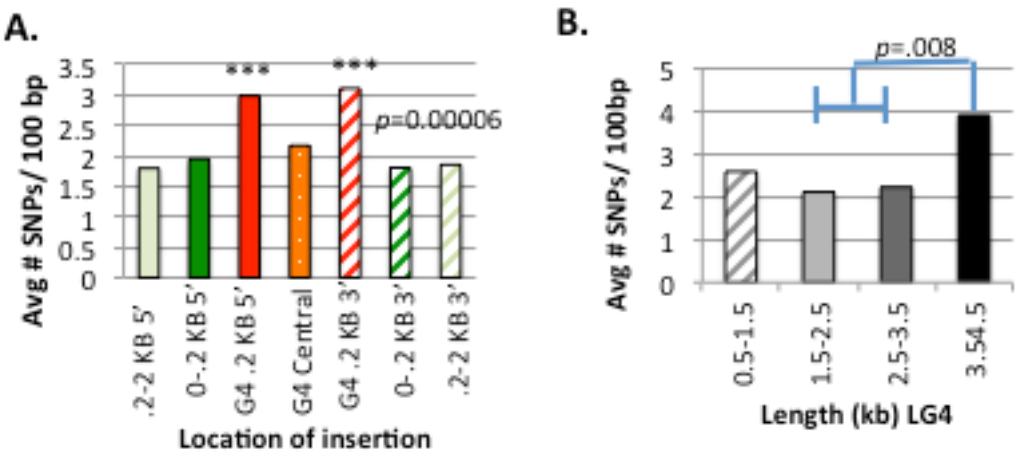

Figure 25. Increase of SNP Density Observed in LG4 Transcribed Regions. Entries off of the dbSNP database were graphed according to individual of #SNPs /100 bp (y axis). **(A)** The average number of SNPs /100 bp (y-axis) was graphed by location with respect to the LG4 sequence and include 0.2-2kb 5'-LG4 (solid light green), LG4 0-0.2 kb 5' LG4 (solid dark green), LG4's first 0.2kb 5' (solid red line), central G4 only (orange dotted), LG4s last 0.2 kb 3' (red striped), flanking 0-.2kb 3' LG4 (striped dark green), and 0.2-2kb 3' LG4 (striped light green). (***) denote significant increase in a one way ANOVA analysis **(B)** The average number of SNPs /100 bp (y-axis) was graphed by the LG4 length (kb) and include 0.5-1.5 kb (gray striped), 1.5-2.5 kb (solid light gray), 2.5-3.5 kb (solid dark gray), and 3.5-4.5 kb (black). The longest LG4s were compared to shorter 1.5-3.5 kb motifs using an unpaired *t*-test.

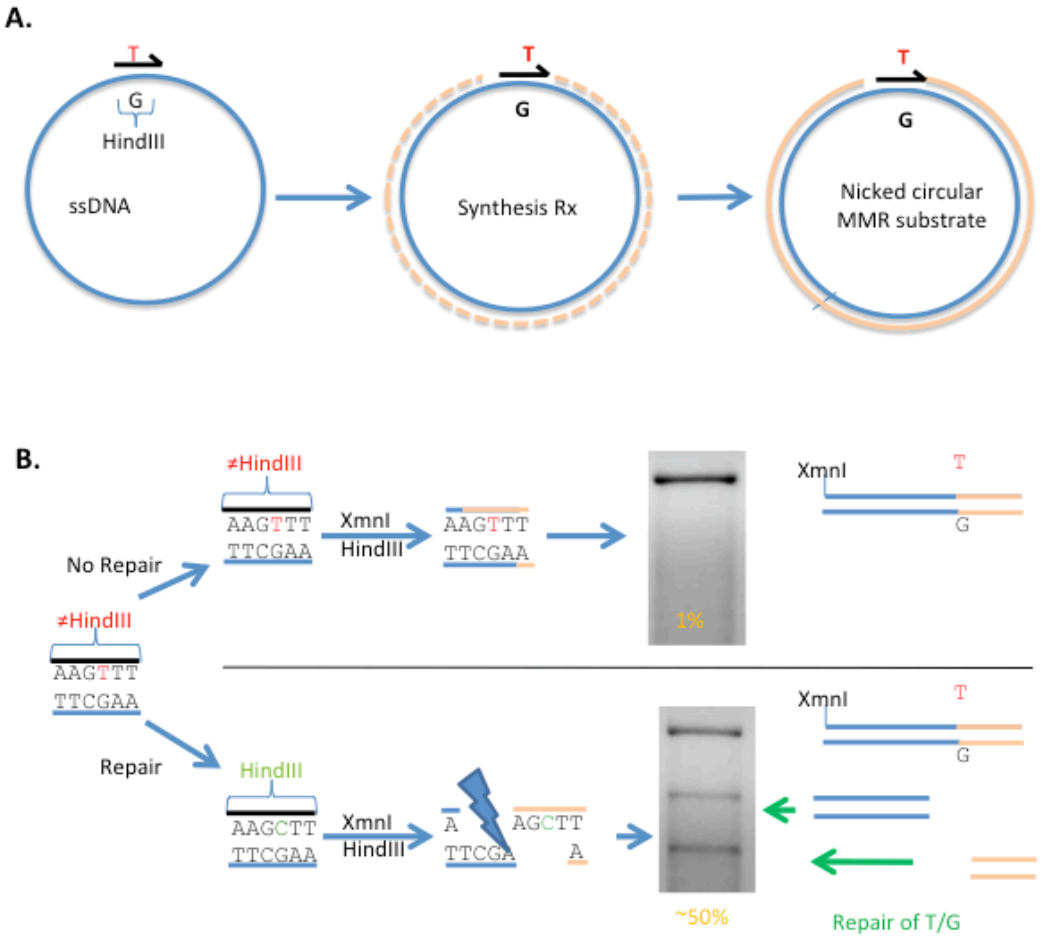

Figure 26. Overview of MMR Substrate Synthesis and Potential *In Vitro* Repair Reaction Outcomes. Depictions of how MMR substrates are synthesized and assayed for repair. **(A)** A primer is annealed to single-stranded DNA (ssDNA) template creating a G-T mismatch at a *Hind*III site, which will inhibit subsequent endonuclease activity unless repaired. Substrate is then produced by a primer extension reaction using the high fidelity Phusion polymerase and purified. (B). Two outcomes of in vitro MMR reactions. No repair of G-T mismatches and subsequent digestion of *Hind*III (incapable) and an adjacent cutter allow only a linearized product or "non-repaired" band (top). If the *Hind*III G-T mismatch is repaired restoring site integrity, double digestion with *Hind*III plus an adjacent cutter and resolution on agarose gel. This allows the detection and measurement of "non-repair bands" verses "repaired bands" (green arrows bottom) by calculation of each bands pixel value on ImageJ. Max repair for control substrates devoid of G4 is ~40-50% depending on freshness of extract prep.

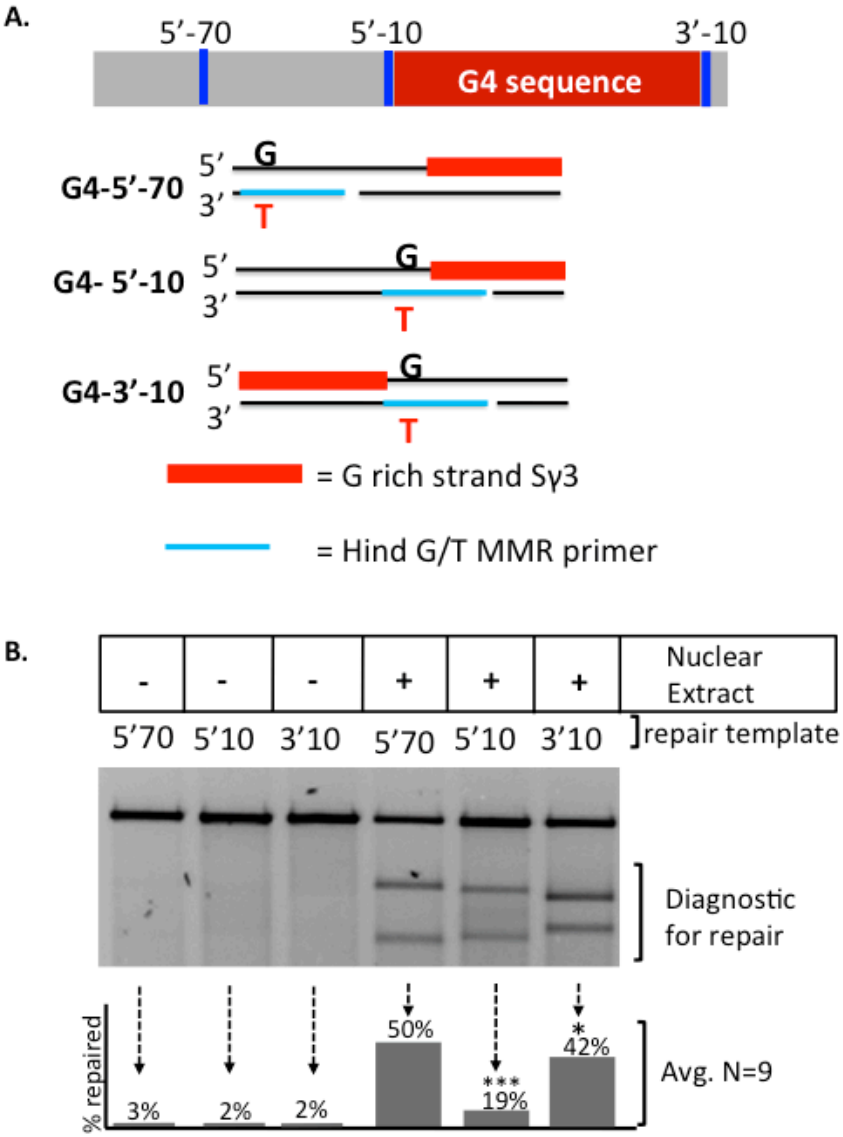

Figure 27. The Repair of G-T Mismatches is Reduced When Directly Next to Sγ3. Ramos extract MMR assays of G-T mismatches at various distances and locations with respect to the G4 capable sequence Sγ3. **(A)** Illustration of substrates used in subsequent MMR assays. All nicks are on the non-G rich strand where repair intermediates could potentially liberate the G4 capable sequence enabling structure formation. The G-T mismatch is 70 bp 5' of Sγ3 (5'-70) and has repair levels similar to wild type controls (not shown). Experimental substrates contained a mismatch 10 bp 5' Sγ3 (5'-10) and 10 bp 3' Sγ3 (3'-10). Location of primer (blue line) with respect to the Sγ3 G-rich sequence (red line) for each substrate is shown. **(B)** An example of an agarose gel of mismatch repair reactions is shown and averages of 9 reactions are graphed below each lane. *P*-values are from two-tailed unpaired t-test (***=$p<.01$, *=$p<.05$).

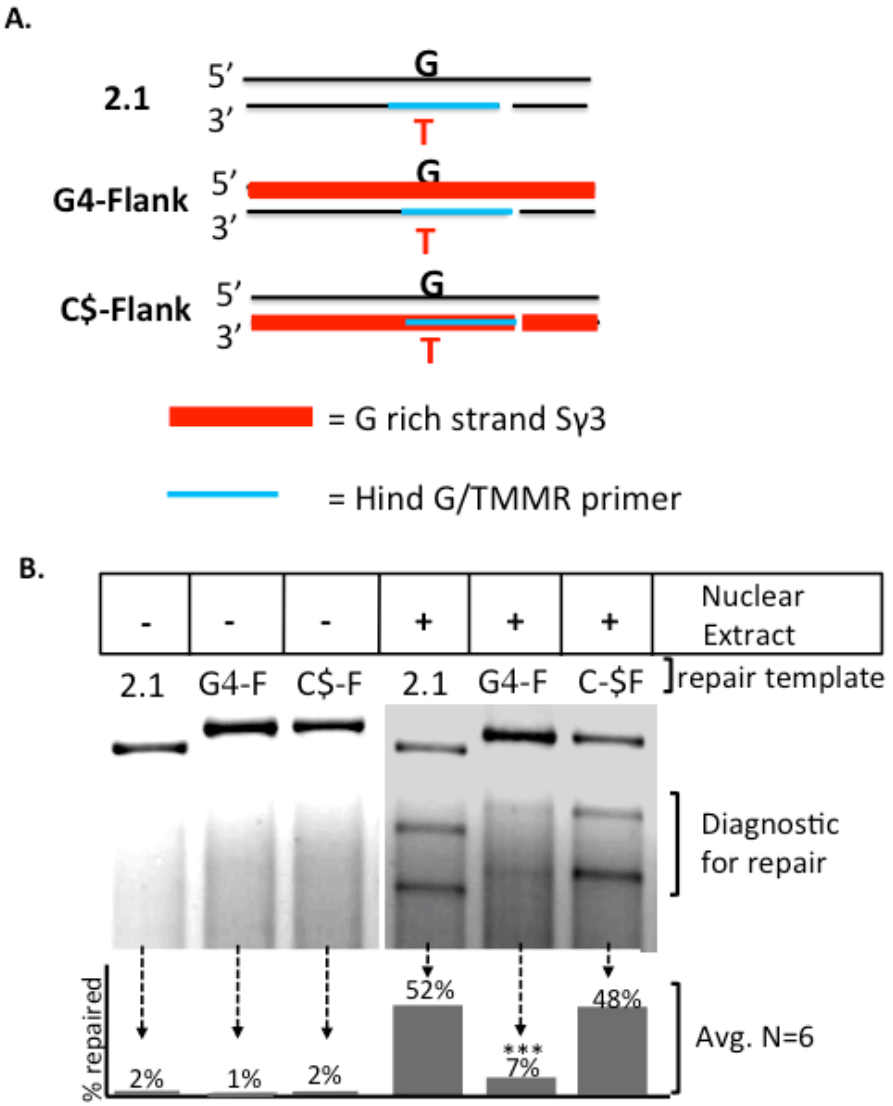

Figure 28. The Repair of G-T Mismatches is Further Reduced in a G4 Orientation Dependent Manner When Directly Inside Sγ3. Ramos extract MMR assays of G-T mismatches in sequences devoid of G4, or in opposite orientation with respect to the G4 capable sequence Sγ3. **(A)** Illustration of substrates used in subsequent MMR assays. The nick is either in a substrate devoid of G4 structures (2.1), on the G-rich strand (C$-Flank), or the non-G-rich strand (G4-Flank). G4-flank is the only substrate where nick directed repair intermediates could liberate the G4 capable sequence enabling structure formation. Location of primer (blue line) with respect to the Sγ3 G-rich sequence (red line) for each substrate is shown. **(B)** An example of an agarose gel of mismatch repair reactions is shown and averages of 6 reactions are graphed below each lane. *P*-values are from two-tailed unpaired t-test (***=$p<.01$, *=$p<.05$).

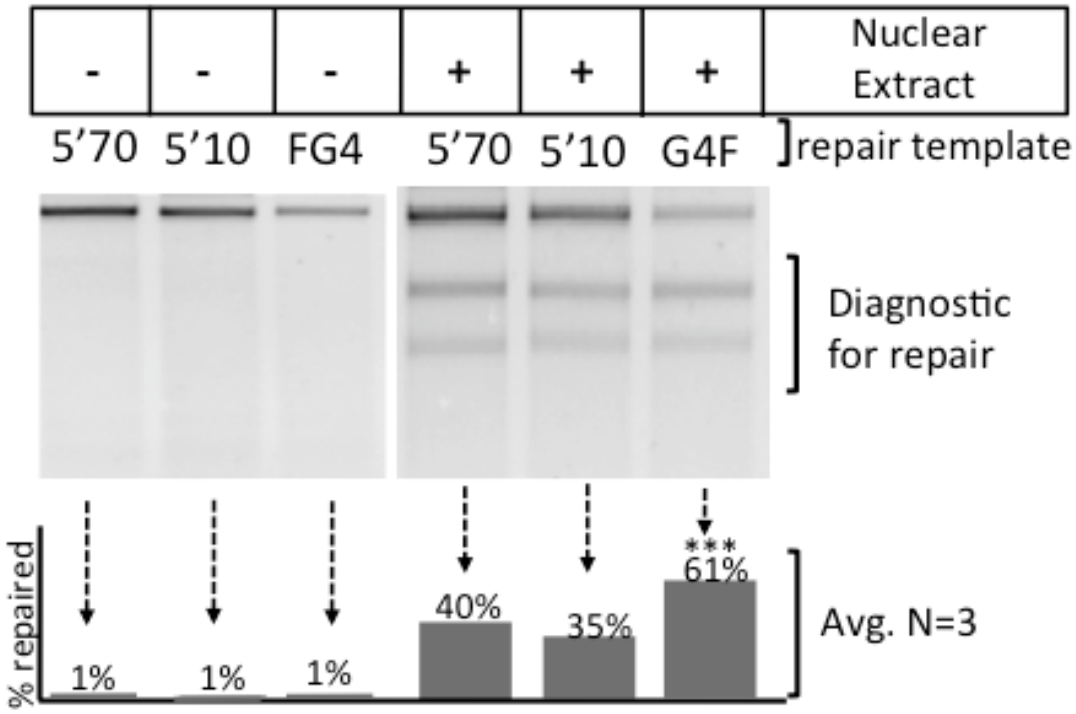

Figure 29. The Repair of Sγ3 G-T Mismatches Are Increased Above Control Substrate Repair in Fresh Repair Reaction Conditions. Substrates used in previous reactions were assayed in fresh repair reaction buffer with an increase of repair levels contradicting previous assays. The exact same Ramos extract was used as previously, suggesting fresh 10X reaction buffer was responsible. Repair in 5'70 stayed at similar levels as before. However, there was a difference in repair in the 5'10 substrate G4-F then previous assays. *P*-values are from two-tailed unpaired t-test (***=$p<.01$, *=$p<.05$).

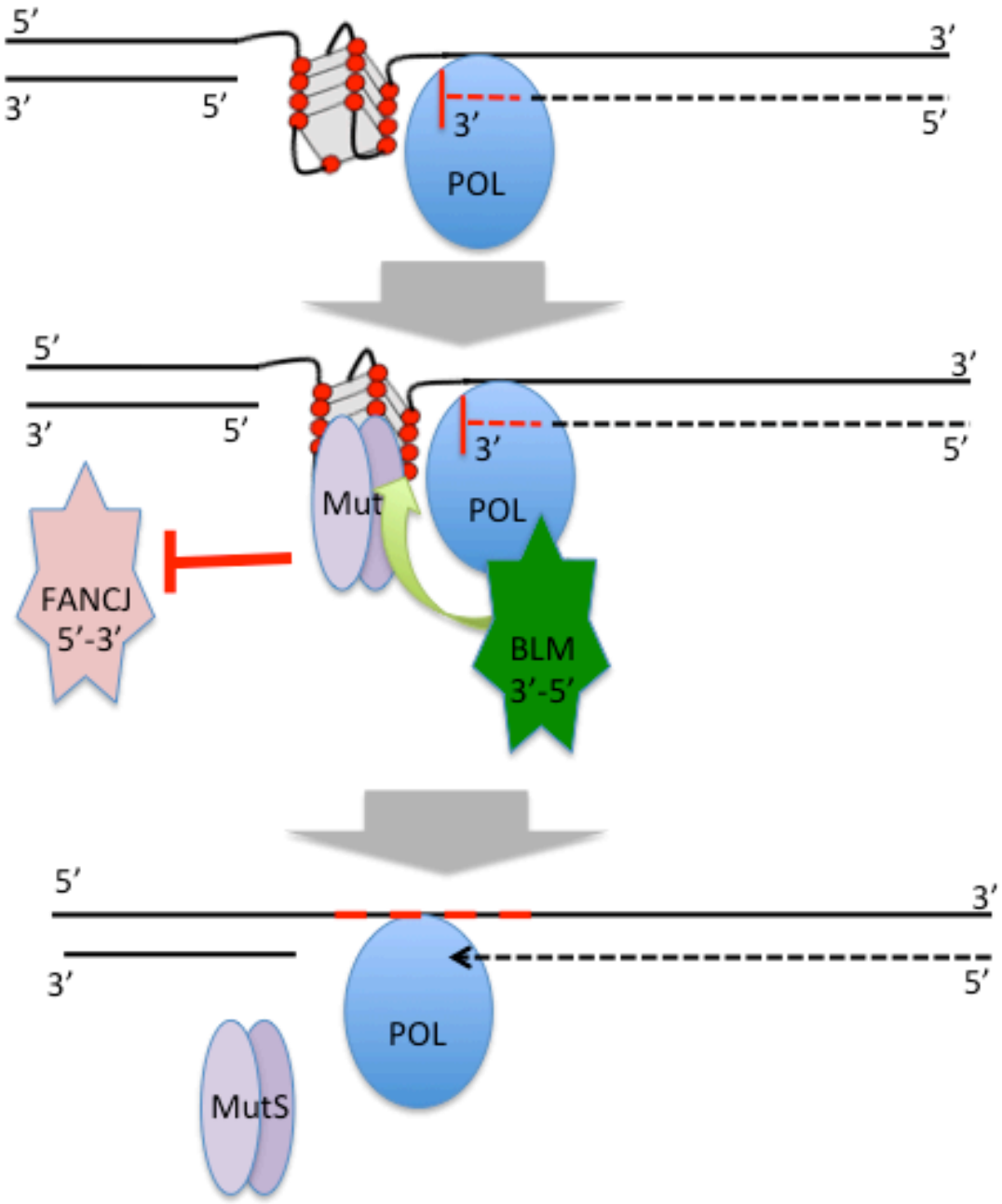

Figure 30. Possible Model for MutS' Role in G4 Structure Resolution. MutS binds to G4 and results indicate a role in structure resolution. I propose a model where MutS binds to G4 structures and directs proper helicase activity to ensure timely and accurate replication through metabolism of G4 structures.

CHAPTER V

DISCUSSION AND FUTURE DIRECTIONS

Guanine (G) rich repeats in the human genome are prone to site-specific genetic instability. This is most likely from the repeats ability to form a stable DNA structure called G-quadruplex (G4). Sequences that support G4 have been computationally identified at 70% of the translocation breakpoints and at multiple oncogenes (Katapadi et al., 2012). However, only a few of these loci have been documented to form G4 structures *in vitro* or *in vivo*. Recent *in vivo* studies have demonstrated that G4 sequences can lead to gross chromosomal rearrangements (GCR) in yeast. The G-rich sequences from these studies were model G4 motifs from human minisatellite (CEB1) (Piazza et al., 2012), and the murine switch region (Smu) (Yadav et al., 2014). To date three G4 motifs found at translocation breakpoints in cancer have been documented to form G4 *in vitro* (Nambiar et al., 2011, 2013; Siddiqui-Jain et al., 2002). Only one of these studies, c-MYC, documented G4's ability to cause GCR (Siddiqui-Jain et al., 2002), while the other analyses focused on G4's biological significance (Nambiar et al., 2011, 2013). These studies have been foundational, but the scope of G4's contributions to genome instability has not been defined. We have yet to fully characterize biologically important G4 loci. Further, the ability of computationally identified G4 motifs to actually form structures *in vitro*, and capability to inflict site-specific genetic instability remains to be determined for many loci. My dissertation provided evidence supporting the model that G4 structures form in the cell from guanine-rich sequences, and that has an important influence on disease, particularly oncogenesis.

Previous analysis has demonstrated that the larger number of consecutive G4 motifs present, the greater the frequency of GCR (Piazza et al., 2012). It was also demonstrated that transcription increases GCR at G4 sequences (Yadav et al., 2014). Immunoglobulin switch regions best exemplify genetic instability triggered by large transcribed G4 sequences. Switch regions are 2-10 kb, highly transcribed introns that participate in programmed recombination (Honjo et al., 2012), and are occasionally sites for chromosomal translocations (González et al., 2007). Therefore, one can predict that other large G4 introns in the human genome may be prone to increased mutagenesis. The ability to identify large G4 sequences was previously limited due to available computational methods, which identified G4 only by looking at individual motifs (Huppert and Balasubramanian, 2005; Todd et al., 2005). We employed a different search program, which returns sites containing large G-triplet dense regions. Over 150 of the largest transcribed G4 motifs (LG4) in the human genome were identified and characterized (Chapter 2). These regions were previously undocumented, and were found in proto-oncogenes, as well as proteins involved in neurological and developmental disease. Further analysis on sequence variation databases provided support that these regions were subject to sequence variations. This study adds to a growing body of evidence that G-rich repetitive sequences promote genome instability. In addition, specific regions of the human genome involved in disease are unstable, and this coincides with sequences that can form G4 structures.

While multiple studies have identified G4 motifs in the human genome, few have investigated the ability of those sequences to actually adopt G4

conformations. Using molecular assays, I tested a subset (10%) of these loci and found that all form G4 structures *in vitro* (Chapter 2). In Chapter 3, I present data of G4-induced instability in *TCF3*, and hairpin formation at a G4 motif (T-5') may play some additional role. In a similar fashion, *HCN2* G4 motifs also contain a similar capability to form hairpins. Interestingly, CD spectra and primer extension reactions indicated that T-5' hairpin formation can interfere with G4 structure formation *in vitro*. However, it is not known if the capacity to form both G4 and hairpin structures is involved in inducing instability, or is more stable than other nearby G4 structures. In summary, beyond simple G4, the ability of a given sequence to adopt additional non-B form conformations should be examined, at least by prediction programs. This may have an additive affect on site-specific instability because other non-B form structures have been found at gross chromosomal rearrangement hotspots (Bacolla et al., 2006).

While it has been known for some time that guanine repeats support the formation of G4 structures, the instability at such genomic regions has not been fully studied. Results in Chapter 2 demonstrate that G4 motifs are variable between sequences, and this has something to do with where that sequence is located within an mRNA. Further, these characteristics may also impact instability. This information could prove invaluable in future deciphering of why specific G4 motifs are prone to high levels of mutagenesis while other are comparatively more stable. For instance, why do TCF3 translocations occur at one G4 motif while others remain apparently stable?

In a study using *C. elegans* as a model organism, DNA breaks at G4 sequences were subject to improper repair and mutagenesis (Koole et al., 2014). Further, research has previously shown that G4 repeats are more likely to induce mutagenesis during transcription when they are located on the non-transcribed strand (Kim and Jinks-Robertson, 2012; Yadav et al., 2014). I did not find any correlation between the orientation of the G4 forming strand in the mRNA and the density of sequence variations, guanine compositions, mRNA location, or regulatory ability. Previous yeast gross chromosomal rearrangement (GCR) assays produced results suggesting a very large increase (38 fold) in rearrangements when G4 sequences were on the non-transcribed strand compared to the transcribed strand (Yadav et al., 2014). Using the identical assay, we only found a 3-fold difference between strand orientations, but GCR levels in the G4 non-transcribed strand were similar to previous experiments (Yadav et al., 2014). Copy number variants in LG4 did not show any strand specificity. Previously, a strand orientation increase in GCR was thought to be due to R-loop formation, or when DNA-RNA hybrids form on the C-rich strand (Duquette et al., 2004). To explain these discrepancies, it is possible that only certain G4 sequences are prone to R-loop formation. Alternatively, R-loop formation is not solely responsible for the observed instability at G4 sequences.

It is also feasible that the orientation of the G4 sequence with respect to the transcribed strand can have an effect on the specific type of mutagenesis that occurs. To date, all of the G4 motifs experimentally shown to form G4 at translocation breakpoints (*c-MYC*, *BCL2*, *HOX11*, *TCF3*) have been in the G-rich

non-transcribed strand orientation (Siddiqui-Jain et al., 2002; Nambiar et al., 2013; Dai et al., 2006; Williams et al., 2015). *TCF3* is an excellent model to study the effect of G4 mRNA strand orientation on genome instability because it contains two large G4 introns on opposite strands. *TCF3*'s t(1:19) breakpoints coincide with G4 structure formation on the non-transcribed strand, while a large increase of CNVs coincide with G4 structure formation on the transcribed strand. In essence, there are two different types of instability observed at these two G4 forming introns. It is possible that translocations do occur at the CNV hotspot (TCF3-lg), but they may not cause disease and so they do not show up in the database. Regardless, I favor a model where G4 orientation, with respect to transcription, impacts the repair pathway choice in response to DNA breaks. In human cultured cells, single-stranded breaks occurring on the transcribed strand increased homology directed repair activities and reduced the more deletion prone non-homologous end joining repair pathway (which acts on non-transcribed strand breaks) (Davis and Maizels, 2014). The potential mutation signature left by inhibition of proper Homology Directed or Non Homologous End Joining repair of breaks (i.e. translocations or indels respectively), coincides with mutagenesis reported around *TCF3*'s G4 regions (Chapter 3). Therefore it is feasible that the G-rich strand orientation with respect to transcription can influence the type of mutagenesis initiated.

Regardless of how G4 DNA induces site-specific instability at guanine-rich motifs, GCR at *TCF3* break point sequences are probably instigated in part by G4 sequences *in vivo*. Considering that mutagenesis of *TCF3* is responsible for

multiple types of leukemia (Hunger, 1996; Aspland et al., 2001; Pui et al., 2004; Schmitz et al., 2012; Steininger et al., 2011), and is frequently translocated in non-small cell lung cancer (Mo et al., 2013), my identification of G4 DNA at those sites helps to explain why the gene is unstable. Bearing in mind that G4 structures have been found at other oncogenes in the human genome (Chapter 2) (Brooks et al., 2010), the effect of G4 on inducing human disease could have implications outside of *TCF3*. Our informatics search alone identified over 60 proteins that can form G4 and are involved in the causation or progression of cancer, neurological, and developmental diseases. While the specific type of mutagenesis occurring in these genes is mostly unknown, investigating how G4 sequences are involved will lay the molecular foundation necessary for clarifying disease etiology.

DNA repair pathways are crucial for maintaining genomic stability. Ironically, DNA repair proteins are also involved in programmed recombination of switch regions (Honjo et al., 2002) and in trinucleotide repeat expansion (McMurray, 2010). Both MutS homologs bind to DNA structures, most likely outside of canonical MMR (Owen et al., 2005; Larson et al., 2005). Further, results above indicate a functional role for MutS in DNA structure resolution (Chapter 2). However, it is unknown what effects MutS-G4 interactions have on genome instability. For example, MutS binds to G4 during programmed recombination (Larson et al., 2005), and could be a factor in inducing off target genomic rearrangements between guanine-rich repetitive loci, such as switch regions and proto-oncogenes. Alternatively, BLM helicase activity in telomere

maintenance is crucial in preventing chromosome aberrations (Barefield and Karlseder, 2012). Therefore, it is possible that MutS-G4 binding partakes in this pathway, and my phage data supports that model. It will require additional experimentation to decipher the complex cellular responses to G4 DNA and how that impacts instability. Clearly, DNA repair activities can be altered or inhibited in G4 sequences (Zhou et al., 2013; Koole et al., 2014). The fidelity of mismatch repair has not previously been assayed in G-rich sequences capable of supporting G4 structures. Results from *in vitro* MMR assays in G4 seem at first glance to be contradicting. However, one consistency in all repair reactions is that when G4 was capable of forming after strand excision, repair levels did not match control levels. This indicates that repair can be inhibited, or increased in G4 sequences. Further clarification on conditions that improve, or inhibit mismatch repair in G4 could provide insights into the mechanism of G4 mediated genome instability, as well as control of programmed recombination.

Although my computational analysis of LG4 transcripts was extensive, there are further analyses that could produce valuable information. For instance, what are the levels of sequence variation in LG4s outside of transcribed regions? This could provide insight on the role transcription plays in inducing genome instability at regions capable of non-B form structures. In addition, one important characteristic of LG4 transcribed regions that were overlooked in my analysis was the length and sequence composition of the loops, or nucleotides in-between the G-repeats. Previous analysis has shown that small looped G4 motifs form more stable structures (Burge et al., 2006). The *TCF3* G4 motif used in

GCR assays contains multiple single and double nucleotide loops (T-3', Chapter 3), so this may suggest that a short loop length is an important factor in instigating genetic instability.

There are most likely multiple overlapping factors that contribute to the induction of mutagenesis at G4 sequences. Copy number variants and indels (Chapter 2) are most likely not the product of MMR. In fact, there is a good probability that the increase of SNPs in LG4s is not completely from inhibition of the MMR pathway (Chapter V). Multiple translesion polymerases are necessary for proper genome maintenance of G4 sequences, and these are typically error prone (Betous et al., 2009; Northam et al, 2014). Considering translesion polymerases have been involved in break repair at G4 (Koole et al., 2014), it is likely that they are active and mutagenic in other repair pathways, including MMR. This has been documented previously in somatic hypermutation where Pol eta is used in conjunction with MMR proteins (Delbos et al., 2005).

The research described here provides support for the hypothesis that G-rich repetitive DNA instigates site-specific genetic instability. Importantly, the impact of that instability may be larger than previously recognized. The ability of these sequences to induce genetic changes is most likely from their ability to form stable four-stranded G4 structures. The largest G4 sequences in the human genome were shown here to from very stable G4 structures *in vitro* (Figure 16). One of them, found at an oncogenic translocation site in *TCF3*, increase chromosomal rearrangements *in vivo*. These results directly connect G4 DNA as a causative contributor to DNA recombination, and mutagenesis. The precise

mechanisms leading to G4-instability are not clear, but some of my research addressed this knowledge gap by examining the role of MMR. My results demonstrate that MMR can be inhibited in an orientation and location dependent manner with respect to the G-rich sequence. While not fully conclusive, my data helps clarify the ability for G4 DNA to instigate mutagenesis in the human genome. Together, the outcomes of my dissertation projects suggest that G4 promotes site-specific instability, and that instability is connected to multiple diseases at a level higher than previously appreciated. It is now clear that the biological impact of G4 sequences on genome instability was previously underestimated. Insights provided here exemplify the depth and complexity of G-rich DNA repeats the human genome and how they may influence genome instability by adopting G4 conformations.

References

Aspland, S. E., Bendall, H. H., and Murre, C. (2001). The role of E2A-*PBX1* in leukemogenesis. *Oncogene, 20*, 5708-5717.

Bacolla, A., Wojciechowska, M., Kosmider, B., Larson, J. E., & Wells, R. D. (2006). The involvement of non-B DNA structures in gross chromosomal rearrangements. *DNA repair*, *5*(9), 1161-1170.

Barefield, C., & Karlseder, J. (2012). The BLM helicase contributes to telomere maintenance through processing of late-replicating intermediate structures. *Nucleic Acids Res*earch, 40(15), 7358-7367.

Bétous, R., Rey, L., Wang, G., Pillaire, M. J., Puget, N., Selves, J., ... & Hoffmann, J. S. (2009). Role of TLS DNA polymerases eta and kappa in processing naturally occurring structured DNA in human cells. *Molecular carcinogenesis*, *48*(4), 369-378.

Brooks, T. A., Kendrick, S., and Hurley, L. (2010). Making sense of G-quadruplex and i-motif functions in oncogene promoters. *Febs Journal*, *277*(17), 3459-3469.

Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K., and Neidle, S. (2006). Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Research*, *34*(19), 5402-5415.

Dai, J., Chen, D., Jones, R.A., Hurley, L.H., Yang, D. (2006). NMR solution structure of the major G-quadruplex structure formed in the human BCL2 promoter region. *Nucleic Acids Research* 34:5133-5144.

Davis, L., Maizels, N. (2014). Homology-directed repair of DNA nicks via pathways distinct from canonical double-strand break repair. *PNAS* 111: E924-E932.

Delbos, F., De Smet, A., Faili, A., Aoufouchi, S., Weill, J. C., & Reynaud, C. A. (2005). Contribution of DNA polymerase η to immunoglobulin gene hypermutation in the mouse. *The Journal of experimental medicine*, *201*(8), 1191-1196.

González, D., van der Burg, M., García-Sanz, R., Fenton, J. A., Langerak, A. W., González, M., Morgan, G. J. (2007). Immunoglobulin gene rearrangements and the pathogenesis of multiple myeloma. *Blood*, 110(9), 3112-3121.

Haeusler, A.R., Donnelly, C.J., Periz, G., Simko, E.A,. Shaw, P.G., Kim, M.S., Wang, J. (2014). C9orf72 nucleotide repeat structures initiate molecular cascades of disease. *Nature* 507:195-200.

Hunger, S. P. (1996). Chromosomal translocations involving the E2A gene in acute lymphoblastic leukemia: clinical features and molecular pathogenesis. *Blood*, *87*(4), 1211-1224.

Huppert, J.L., Balasubramanian, S. (2005). Prevalence of quadruplexes in the human genome. *Nucleic Acids Res*earch 33:2908–2916.

Honjo, T., Kinoshita, K., & Muramatsu, M. (2002). Molecular mechanism of class switch recombination: linkage with somatic hypermutation. *Annual review of immunology,* 20(1), 165-196.

Larson, E. D., Duquette, M. L., Cummings, W. J., Streiff, R. J., & Maizels, N. (2005). MutSα binds to and promotes synapsis of transcriptionally activated immunoglobulin switch regions. *Current biology*, *15*(5), 470-474.

Katapadi, V. K., Nambiar, M., & Raghavan, S. C. (2012). Potential G-quadruplex formation at breakpoint regions of chromosomal translocations in cancer may explain their fragility. *Genomics*, 100(2), 72-80.

Kim, N., Jinks-Robertson, S. (2012). Transcription as a source of genome instability. *Nature* Reviews Genetics 13:204-214.

Koole, W., van Schendel, R., Karambelas, A. E., van Heteren, J. T., Okihara, K. L., & Tijsterman, M. (2014). A Polymerase Theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites. *Nature* communications 5.

Maizels, N. (2012). G4 motifs in human genes. *Annals of the New York Academy of Sciences* 1267: 53-60.

McMurray, C.T. (2010). Mechanisms of trinucleotide repeat instability during human development. *Nature* Reviews Genetics 11:786-799.

Mo, M. L., Chen, Z., Zhou, H. M., Li, H., Hirata, T., Jablons, D. M., and He, B. (2013). Detection of E2A-*PBX1* fusion transcripts in human non-small-cell lung cancer. *J Exp Clin Cancer Res*, *32*, 29.

Nambiar, M., Goldsmith, G., Moorthy, B. T., Lieber, M. R., Joshi, M. V., Choudhary, B., Raghavan, S. C. (2011). Formation of a G-quadruplex at the BCL2 major breakpoint region of the t (14; 18) translocation in follicular lymphoma. *Nucleic Acids Res*earch 39:936-948.

Nambiar, M., Srivastava, M., Gopalakrishnan, V., Sankaran, S. K., & Raghavan, S. C. (2013). G-quadruplex structures formed at the HOX11 breakpoint region contribute to its fragility during t (10; 14) translocation in T-cell leukemia. *Molecular and cellular biology*, 33(21), 4266-4281.

Northam, M. R., Moore, E. A., Mertz, T. M., Binz, S. K., Stith, C. M., Stepchenkova, E. I., Shcherbakova, P. V. (2014). DNA polymerases ζ and Rev1 mediate error-prone bypass of non-B DNA structures. *Nucleic Acids Res*earch, 42(1), 290-306.

Owen, B. A., Yang, Z., Lai, M., Gajek, M., Badger, J. D., Hayes, J. J., McMurray, C. T. (2005). (CAG) n-hairpin DNA binds to Msh2–Msh3 and changes properties of mismatch recognition. *Nature structural & molecular biology*, *12*(8), 663-670.

Pui, C. H., Relling, M. V., and Downing, J. R. (2004). Acute lymphoblastic leukemia. *New England Journal of Medicine*, *350*(15), 1535-1548.

Piazza, A., Serero, A., Boule, J. B., Legoix-Ne, P., Lopes, J., and Nicolas, A. (2012). Stimulation of gross chromosomal rearrangements by the human CEB1 and CEB25 minisatellites in Saccharomyces cerevisiae depends on G-quadruplexes or Cdc13. *PLoS genetics*, *8*(11), e1003033.

Schmitz, R., Young, R. M., Ceribelli, M., Jhavar, S., Xiao, W., Zhang, M., Wright, G., Staudt, L. M. (2012). Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature*, *490*(7418), 116-120.

Siddiqui-Jain, A., Grand, C. L., Bearss, D. J., & Hurley, L. H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proceedings of the National Academy of Sciences* 99:11593-11598.

Steininger, A., Möbs, M., Ullmann, R., Köchert, K., Kreher, S., Lamprecht, B., Anagnostopoulos, I., Assaf, C. (2011). Genomic loss of the putative tumor suppressor gene E2A in human lymphoma. *The Journal of experimental medicine*, *208*(8), 1585-1593.

Tarsounas, M., Tijsterman, M. (2013). Genomes and G-quadruplexes: for better or for worse. *Journal of molecular biology* 425:4782-4789.

Todd, A.K., Johnston, M., Neidle, S. (2005). Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res* 33:2901–2907.

Williams, J.D., Fleetwood, S., Berroyer, A., Kim, N., Larson, E.D. (2015). F Formation of G-quadruplex DNA influences the genetic stability of human TCF3 (E2A). *Frontiers Journal* Submitted February

Yadav, P., Harcy, V., Argueso, J.L., Dominska, M., Jinks-Robertson, S., Kim,. N (2014). Topoisomerase I Plays a Critical Role in Suppressing Genome Instability at a Highly Transcribed G-Quadruplex-Forming Sequence. *PLoS genetics* 10:e1004839.

Zhang, Y., Yuan, F., Presnell, S.R., Tian, K., Gao, Y., Tomkinson, A.E., Li, G.M. (2005). Reconstitution of 5′-directed human mismatch repair in a purified system Cell 122: 693-705.

Zhou, J., Liu, M., Fleming, A. M., Burrows, C. J., & Wallace, S. S. (2013). Neil3 and NEIL1 DNA glycosylases remove oxidative damages from quadruplex DNA and exhibit preferences for lesions in the telomeric sequence context. Journal of Biological Chemistry, 288(38), 27263-27272.